Online computer-based testing in human resource management

Schakel, L

Online Computer-based Testing in Human Resource Management: Contributions from Item Response Theory

Lolle Schakel

The research presented in this dissertation was supported by PiCompany BV. The opinions expressed by authors are their own and do not necessarily reflect the views of PiCompany BV.

ISBN: 978-90-367-5821-5
ISBN electronic version: 978-90-367-5820-8
Cover designed by Bruksvoort Design & Content
Layout by Lolle Schakel
Printed by Ipskamp Drukkers B.V., Enschede

RIJKSUNIVERSITEIT GRONINGEN

Online Computer-based Testing in Human Resource Management: Contributions from Item Response Theory

Proefschrift

ter verkrijging van het doctoraat in de
Gedrags- en Maatschappijwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
donderdag 6 december 2012
om 14.30 uur

door

Lolle Schakel

geboren op 23 mei 1960
te Wonseradeel

Promotor:	Prof. dr. R.R. Meijer
Copromotor:	Dr. I.J.L. Egberink
Beoordelingscommissie:	Prof. dr. H.A. Hoekstra
	Prof. dr. K. Sanders
	Prof. dr. K. Sijtsma

Contents

Chapter 1
New Developments in the Use of Personality Questionnaires in HRM

1.1 Introduction

Cognitive ability tests and personality questionnaires are frequently used to predict future job success. They are used to select people for new jobs, to support individuals wishing to move on to another job and in career pathways, and as such they are indispensable tools for Human Resource Management (HRM) staff.

Test administration is coupled with the further automation of work processes within HRM. Personnel data is administered via systems that offer scope for recording an individual's training, qualifications, skill levels and appraisals, alongside personal information. Information of this kind can be used to track and guide an employee's personal development within an organization. To an increasing extent, much of the selection process is being managed online, via computers and the Internet, with applicants asked to lodge their résumé and other relevant information via a website. This gives recruiters and psychologists access to a range of data even before they have any personal contact with applicants, enabling them to select the best candidates in advance.

The option of applying for jobs online has advantages for applicants too. They can undergo the first part of the selection process at a time and place that suits them. Applicants can apply outside work time and do not have to take time off for the first interview until their résumé and psychological test data show that they are indeed suitable candidates for the job. Online-based applications make it easy for applicants to apply for many vacancies at the same time, as well as with different organizations.

These new ways of applying for jobs make it necessary for employers to communicate with potential candidates in new ways. In their online selection process, employers have to set themselves apart from their competitors and remove any obstacles confronting applicants. In particular, the shortage of highly qualified and talented staff in the job market (Bersin, 2011; Guthridge, Komm & Lawson, 2008) makes it essential to recruit the best candidates via online selection as well.

This means that the online selection process has to be clear and transparent. It also should not be time consuming, potentially causing candidates to stop partway through.

Test and questionnaire developers have to anticipate these new trends and the changing role that measurement instruments will play. Increasingly, tests and questionnaires are delivered in an unproctored setting, which means that there is no supervisor present during the test administration. Online testing requires short questionnaires, because of the short attention span of Internet users. Also, the purpose for which test scores and questionnaire results are used is becoming blurred, as information gathered during selection is entered into HRM systems and can be used at a later date for an employee's career development (Burke, 2011). Once test information is stored in databases, it can be used many times over to match an individual with various jobs. This procedure demands that test information should be handled carefully and that participants should be clearly informed about how the data will be used. In the Netherlands, test information and its use falls under legislation governing the protection of personal information. The Dutch Data Protection Authority (CBP, www.cbpweb.nl) ensures that personal information is properly used and secured in order to safeguard the privacy of individuals today and in the future.

In computer-based testing item response theory (IRT, Embretson & Reise, 2000) presents test developers interesting tools to evaluate and construct new instruments (e.g., Egberink & Meijer, 2012). In this thesis various studies are presented in which IRT plays a key role in the development and evaluation of computer-based and online measurement instruments within HRM. The emphasis is on personality testing. Therefore, before the outline of this thesis is presented, the background and use of personality questionnaires for HRM purposes are discussed.

1.2 The Use of Personality Questionnaires within HRM

The idea that personality plays a role in whether or not someone will perform well in a particular job goes back a long way. It has long been understood that, in addition to mental abilities affecting how people carry out tasks, personality differences could explain their eventual performance. Before he developed the first Binet scale, Binet recognized the impact of personality on intellectual functioning (Binet & Henri, 1895). Although this idea was accepted by many researchers and was met with much support, it was not until the First World War that specific questionnaires about non-

cognitive functioning were developed. This move was prompted by the large number of men in the United States wishing to enlist for military service at that time. A key element in the selection of soldiers was to be able to identify emotionally unstable applicants and to bar them from jobs with the army. Psychiatric interviews, the only available instrument, were insufficient for coping with the flood of applications. Psychologists therefore developed a self-assessment questionnaire, the Woodworth Personal Data Sheet, to gauge the mental stability of potential soldiers and to make the selection process more efficient (Kaplan & Saccuzzo, 2005).

The use of personality measures for selection decisions has burgeoned since the 1990s. Two meta-analyses have shown that, together with intelligence measures, personality measures have an added value for predicting future job success (Barrick & Mount, 1991; Tett, Jackson & Rothstein, 1991).

Over the years, a lot of personality questionnaires has been developed. Some well-known examples are the Minnesota Multiphasic Personality Inventory (MMPI: Hathaway & McKinley, 1943, Ben-Porath & Tellegen, 2008), the Jackson Personality Inventory (JPI: Jackson, 1976, 1997) and the Myers-Briggs Type Indicator (MBTI: Myers, 1962). An often used questionnaire is the NEO Personality Inventory-Revised (NEO-PI-R; Costa & McCrae, 1992, for the Dutch version, see Hoekstra, Ormel & de Fruyt, 2007). This questionnaire is based on the Five Factor Model (FFM), also called the 'Big Five', which has become the most important personality model in psychology. These questionnaires have a long research tradition and are published by reputable test publishers.

In addition to these questionnaires, which have been subjected to scientific study, there are numerous popular personality tests, most of them are short. Because people like to explore who they are and discuss this with others, personality tests of this kind regularly feature in lifestyle and other popular magazines (such as Quest Psychologie, 2011). The Internet, Facebook, and other social media offer a host of free questionnaires for anyone to complete. A Google search for free personality tests in January 2011 yielded more than 13,000 hits in Dutch and almost eight million in English. Assessing the psychometric quality of these tests can be difficult, as often the only reference points are their layout and the clarity with which results are interpreted. It is therefore difficult for a layperson to assess the value of the results of this type of questionnaires.

The BBC TV programme 'Child of Our Time' has made personality testing based on the Big Five accessible to a large group of people in the UK. Millions have taken part in the free personality questionnaire (www.bbc.co.uk/labuk/experiments/personality), which was offered in connection

with the programme. Various personality tests can also be taken, either free of charge or for a small fee, via websites such as 123test.nl. Thus the concept of personality is now familiar and accessible to many persons. As a consequence, it is common for job applicants to have already completed a personality questionnaire and to be aware of the relationship between their answers and the results, before they fill out a questionnaire as part of the selection procedure. For test publishers this means that their personality questionnaires, which are often better substantiated and researched, must compete with free tests that tend to be less well researched.

1.3 Validity of Personality Questionnaires within HRM

A great deal of research has been done on the predictive value of various psychological instruments. A meta-analysis by Schmidt and Hunter (1998) shows that cognitive ability tests have a predictive validity for work performance of about .50 on average, while Conscientiousness questionnaires have a predictive validity of about .30. Schmidt and Hunter (1998) also describe which combination of measurement instruments produces the best prediction (often referred to as 'incremental validity'). They show that the predictive validity of cognitive ability tests can be increased by adding a personality test, especially a Conscientiousness questionnaires, to the selection procedure. Likewise, the incremental validity of this combination of tests can be increased further if a structured interview is added to the mix (Schmidt & Hunter, 1998). It is generally recommended not to use a personality questionnaire in isolation, but always to follow up with a structured interview. The questionnaire functions as an efficient way to find out about the candidate's general strengths and weaknesses in relation to the job description. The purpose of the interview is to establish personal rapport with candidates and to explore how they handle their strengths and weaknesses in work situations.

There has been much discussion recently about the use of self-report questionnaires as a personality measure for personnel selection (Morgeson et al., 2007a, 2007b). The most important criticisms are the low predictive value for general performance as revealed in new meta-analyses and the ease with which this type of questionnaires can be answered in a socially desirable way. Morgeson et al. (2007a) take these criticisms so seriously that they advise caution about the use of self-questionnaires as a decision criterion in personnel selection. In response to Morgeson et al. (2007a, 2007b), various researchers point out that the Big Five factors are too broad for predicting job success. For example, Tett and Christiansen

(2007) suggest the use of subfactors in combination with personality-based job analysis in order to increase the predictive validity of personality tests.

The use of personality questionnaires for career development is less contested. Here, filling out the questionnaire in a socially desirable way is not in a candidate's interest. The outcomes of the measure are to be matched with suitable jobs and work behaviour which can easily be developed in view of the candidate's personality profile.

1.4 New Developments in the Use of Personality Measures within HRM

1.4.1 Unproctored Computer-Based and Online Testing

Computer-based personality questionnaires are increasingly taking the place of paper-and-pencil versions. This raises the question whether a computer-based administration is equivalent to a paper-and-pencil version. Extensive research on this issue reveals that this is indeed often the case (Chuah, Drasgow & Roberts, 2006; Potosky & Bobko, 1997; Richman, Kiesler, Weisband & Drasgow, 1999; Wilkerson, Nagao & Martin, 2002). Salgado and Moscoso (2003), for example, find the same distribution of total scores and a similar reliability and factor structure for Big Five paper-and-pencil personality questionnaires and those administered via the computer.

Online testing is a logical follow-up to computer-based testing. Whereas with computer-based testing the questionnaire is installed on the computer on which the test is taken, or on an enterprise network, in the case of online testing the test can be taken on any device that has Internet access. This could be a computer terminal at a test venue, but also a personal computer at home, a mobile phone, an information kiosk at an employment agency, or a device in an Internet café or restaurant. In other words, delivery conditions may vary enormously.

Because of the advantages over traditional delivery, online testing is set to play an ever greater role within HRM (Tippins, 2009b). There are significant benefits in terms of cost savings and a more efficient selection process, prompting a growing number of employers to insist that tests should be delivered where possible via the Internet before any face-to-face contact with candidates (Tippins, 2009a). At the same time, Industrial and Organizational psychologists are still debating the admissibility of online tests. Many professional guidelines for psychologists do not

provide guidelines for the use of tests that are not supervised by a psychologist. The Dutch Psychological Association (NIP), for example, does not mention this option in its rating system for test quality (Evers, Sijtsma, Lucassen, & Meijer, 2010) or in its information for test candidates (*Onderzoek bij testkandidaten*, n.d.). Nor do the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) describe any procedures for the use of unproctored tests (Tippin, 2009a, 2009b). Only the International Test Commission (2005) has drawn up guidelines for computer-based and online testing. They make a distinction between an "open mode" and a "controlled mode" (both unproctored) as well as a "supervised mode" and a "managed mode" (both proctored). In the open mode the test taker has direct access to the test materials and there is no involvement of a test user or test administrator. The controlled mode is a mode of test administration in which control is exercised over who can access a test on the internet and how often, where and when they can access it. In the supervised mode the test administrator has direct face-to-face involvement with the test taker. The test takers will come to a location where the test administrator is able to supervise them taking the test. However, the test distributor has no means of directly controlling the nature of the location or the type of equipment being used. A managed mode is a administration in which there is both direct supervision and control over the equipment being used, and other conditions. Typically managed mode administration refers to the use of dedicated testing centres.

Advantages and Disadvantages of Unproctored Online Testing

The advantages of online tests are partly in line with those of computer-based tests (Baron & Austin, 2000) and can be summarized as improving the quality and efficiency of test delivery. Like its computer-based counterpart, online testing ensures that all candidates receive the same instructions and are given exactly the same time in which to answer the questions. The automatic scoring of answers minimizes the risk of incorrect calculations. A further advantage is that any changes and updates can easily be introduced. The biggest advantage of online testing, however, is that the test can be taken anywhere and at any time.

Online tests have different disadvantages depending on their type. There are other objections to the delivery of non-cognitive tests via the Internet than to that of cognitive tests (Arthur, Glaze, Villado & Taylor, 2010; Tippins, 2009a). The chief objection to delivering personality and other non-cognitive tests in this way is the possibility of influencing the responses, quite apart from the issue of whether the

tests are delivered remotely in a proctored or an unproctored setting. The size of this effect remains a subject of discussion (e.g. Ellingson, Sackett & Connelly, 2007; Hogan, Barrett & Hogan, 2007; Hough & Oswald, 2008; Viswesvaran & Ones, 1999). Giving socially desirable answers is a problem in situations where important consequences are attached to the outcome of the questionnaire, such as in job selection procedures. This is less of a problem when the stakes are not so high. Arthur, Glaze, Villado, and Taylor (2010) showed that individuals scored higher on four of the five Big Five personality factors, with the exception of Openness, when the online personality test was part of a selection procedure than they did one year later when completing the same questionnaire for research purposes. Thus the issue of socially desirable scoring will continue to play a significant role in online testing as well.

Using multidimensional forced-choice (MFC) questionnaires is one way of making personality questionnaires more resistant to socially desirable response behaviour (Chernyshenko et al., 2009; Brown & Maydeu-Olivares, 2011). This type of tests presents test-takers with two or more statements of similar desirability, but in different domains. The test-taker must indicate which statement best describes him or her. Recent empirical research (Stark & Chernyshenko, 2011) shows promising results with the use of multidimensional pairwise preference tests, a variant of MFC tests. One disadvantage from a practical point of view is that psychometric methods for producing good forced-choice questionnaires are still under development.

1.4.2 A Call for Shorter Questionnaires

The Internet is a transient environment, where people have a short attention span. This places demands on questionnaires, which must be short and concise as well as reliable (see Emons, Sijtsma & Meijer, 2007 for a critical discussion). The demand for short questionnaires also applies to the fields of education (Sinharay, Puhan & Haberman, 2010) and clinical psychology (Fliege et al., 2009; Gibbons et al., 2008). Here, too, there is a need for reliable measurements of as many concepts as possible using as few items as possible. For educational testing, in addition to the general score in a subject area, assessors are often interested in diagnostic advice. In the case of a math test, for example, in addition to the general score there will be an interest in knowing on which subcomponents the candidate performed best and worst – does the candidate have to do more work on geometry, or is he or she underperforming in algebra? Haberman and Sinharay (2010) point out the dangers of overinterpreting these subscores. Subscores are often unreliable and have a high

correlation with the total score, which means that they contribute little additional information over and above the total score (Puhan, Sinharay, Haberman, & Larkin, 2010).

Many different personality and mood questionnaires have traditionally been used for diagnostic purposes in clinical psychology. Clients often have to complete several lengthy questionnaires, containing hundreds of questions. Abbreviated versions of various questionnaires are available for self-diagnostics and online therapy. In the Netherlands, for example, Interapy (n.d.) uses abbreviated questionnaires for its online therapies. And the PROMIS programme (Choi, Reise, Pilkonis, Hays & Cella, 2010) has shown that short questionnaires containing seven to eight items can effectively measure the concepts of depression, anxiety, and anger. Pilkonis et al. (2011) report α coefficients of .90 through .95. These questionnaires are constructed on the basis of research into the most discriminating items in a large pool of items (or item bank). A point of discussion with regard to these questionnaires is exactly which content the items measure, as many items are shown to ask similar content (Reise & Waller, 2009).

1.4.3 Portals and Database Storage of Test Information

A portal is an online environment where individuals are invited to lodge their résumé and to complete tests and questionnaires for the purpose of job applications or career advice. Within such a portal individuals can gain some idea of their own strengths and weaknesses in relation to the job market and to job vacancies, and the organization in question can obtain information about the capabilities of the participants in relation to potential clients or vacancies. On the basis of psychological constructs such as intelligence, personality, values and motivation, the individual is matched to career options or vacancies.

The provision and sharing of information by the individual in this way can be compared to building up a personal portfolio. In a career development context, this portfolio is akin to one intended for professional certification, containing information on education and/or training and competencies acquired elsewhere. In a selection context, however, the purpose of the portfolio is rather to enable an existing or potential employer to gather as much relevant information as possible about an individual's suitability for a particular job. It is, therefore, vital for portal participants to know what the information will be used for. If it is solely in their own interests, they are more likely to respond as truthfully as possible. However, if the employer's interests weigh more heavily, participants will probably profile themselves in such a way as to maximize their chances of getting a new job or

furthering their career. The risk of giving socially desirable responses to questionnaires is greater in such cases, and reusing such test information for career advice purposes can have a negative impact on the quality of that advice.

1.5 Test Construction Based on Item Response Theory

The modern technological developments described above have resulted in new forms of test construction and delivery. A well-known example is computerized adaptive testing (CAT: Meijer & Nering, 1999; van der Linden & Glas, 2010), which is often based on IRT (Embretson & Reise, 2000). IRT does not focus on the test score, but on the item and the response to that item. IRT explains item scores by postulating a latent trait (often symbolized by the Greek letter θ), such as extraversion and the item characteristics such as the item difficulty. With an IRT model it is possible to place person parameters (θ) and item parameters on the same scale. As a consequence, an individual's position on the latent trait continuum can be estimated from his or her responses to a random subset of items from a large pool of tems (item bank). The use of CAT in combination with IRT has some advantages over testing using fixed questionnaires. For example, Hol, Vorst and Mellenbergh (2005, 2007) showed that CATs for personality require fewer items in order to achieve reliability comparable to that of regular test delivery using a fixed set of items. In cognitive and educational testing there are also some examples of more efficient test use (e.g., Rudner, 2010). In other words, the use of CAT and IRT models enables a researcher to develop short personality questionnaires focusing on a specific topic. And the development of an item bank means that a candidate can be presented with a new set of items at any time.

Other examples of using IRT to improve test quality involve techniques for exploring differential item functioning (Zwick, 1990), the study of invariant item ordering (Ligtvoet, van der Ark, te Marvelde, & Sijtsma, 2010; Meijer & Egberink, in press) and the use of equating and linking procedures (McHorney & Cohen, 2000).

Although, in practice, there are many computer-based developments in both cognitive and non-cognitive testing, the application of IRT to obtain a stronger psychometric basis for the measurement instruments is scarce in the HRM-literature. In this thesis, I use different IRT techniques to obtain a better insight into the psychometric quality of different measurement instruments. On the other hand, these applications provide psychometricians information about the performance of IRT methods under realistic conditions.

1.6 Outline of This Thesis

Before the use of IRT in HRM is illustrated, Chapter 2 describes the development of an online-administered computer-based Big Five instrument for the workplace, the Reflector Big Five Personality (RBFP), based on classical test theory. This is done, because in Chapters 3 through 5 different psychometric methods, based on IRT, are applied to this instrument. The RBFP was developed in a Dutch human resources assessment company, throughout this thesis this company is referred to as the Company. In Chapter 3, IRT-based differential functioning of the RBFP is investigated in two contexts, a selection context and a career development context. In Chapter 3, first, scaling results in both contexts are reported. Second, differential item and test functioning are investigated using a likelihood ratio approach and using different recently proposed effect size measures. Results showed that the scalability was lower in the selection context than in the career development context, but that differential test functioning was of no practical importance. In Chapter 4 the usefulness of CAT for personality in a real life workplace counseling context is investigated. A sample of candidates completed the CAT as part of their career development procedure. Results showed that CAT resulted in a reduction of items administered and administration time, whereas high correlations were found between CAT and full scale scores. However, the item pool was not very suited to discriminate candidates with moderate to high values on the investigated personality traits. Chapter 5 is devoted to the role that invariant item ordering can play when selecting items for short versions of the RBFP. In contrast to the first five chapters, Chapter 6 is devoted to intelligence testing. In Chapter 6, the results on a CAT for intelligence are compared between the unproctored and proctored setting. Results showed that for most candidates scores were similar. For those persons that produce large differences between test scores in the unproctored and proctored setting, a method is proposed through which additional diagnostic information can be obtained.

The chapters in this thesis are self-contained, hence they can be read separately. Therefore, some overlap could not be avoided.

Chapter 2
A Work-Related Personality Questionnaire: The Reflector Big Five Personality

2.1 Introduction

In this chapter the theoretical and psychometrical background of the Reflector Big Five Personality (RBFP) questionnaire is discussed. This is done, because in Chapters 3 through 5 different psychometric methods, based on IRT, are applied to this instrument. In Chapter 3 differential item and test functioning of the RBFP is investigated using different types of effect size measures, in Chapter 4 a computerized adaptive version of the RBFP is developed and discussed, and in Chapter 5 the property of invariant item ordering (IIO) is investigated for the RBFP.

The RBFP is an online-administered computer-based Big Five instrument; therefore the Big Five model and its use within human resource management (HRM) are discussed first. Second, the development of the RBFP and some research studies regarding its psychometric quality are described. Finally, the online administration and reporting process are discussed.

2.2 Big Five Model

The Five Factor Model (FFM) of personality has brought structure to much research into the nature of personality. Before the advent of the FFM, there was little consensus regarding the structure of personality traits. A diversity of instruments and scales were developed, each of which conceptualized personality in its own way (e.g., Gough, 1957).

The FFM originates from the factors found in research into words that describe people in various languages (De Raad, 1992; Digman, 1990; Goldberg, 1990, 1993). In countless publications at the beginning of the 1990s, new evidence was put forward for describing personality by means of five factors. Personality researchers

reached consensus on the idea that the five personality constructs Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience provided adequate to satisfactory descriptions of the basic dimensions of 'normal' personality (see Borkenau & Ostendorf, 1990; Costa & McCrae, 1998; Digman, 1990; Goldberg, 1990, 1993; McCrae & Costa, 1987). Later, a discussion arose as to whether these five factors constitute a sufficiently valid description of personality. Lee and Ashton (2004) advocated a six-factor model, the HEXACO model. In addition to the Big Five factors, they identified the integrity dimension Honesty-Humility. The model also incorporates changes in the positions of the Agreeableness and Emotionality axes, which have been rotated in relation to the Agreeableness and Emotional Stability axes in the Big Five model. Almagor, Tellegen, and Waller (1995) described a model with seven factors (the Big Seven). They identify 2 evaluative factors - Positive and Negative Valence - in addition to the Big Five factors. Saucier (2003) also described seven factors in the 'Multi-Language Seven' (ML7) factor model. Whereas five-factor models comprise three affective–interpersonal factors (Extraversion, Emotional Stability, and Agreeableness), the ML7 has four (Gregariousness, Self-Assurance, Even Temper, and Concern for Others). The ML7 partitions negative emotionality into two factors, one that is more related to fear (low Self-Assurance), the other more to anger and hostility (Temperamentalness versus Even Temper). Other authors propose extra dimensions on the basis of cross-cultural research (e.g. Cheung et al., 2001), although the five factors still occur in studies in various countries and cultures (McCrae & Allik, 2002; Schmitt et al., 2007). The discussion surrounding the number of factors is not unexpected, given that the factors are usually derived from factor analysis of a heterogeneous set of behaviours, feelings, and thoughts. The number of factors found depends on the variables used in the analyses, their psychometric properties, and the extent to which researchers accept heterogeneity in their solution (Chernyshenko, Stark & Drasgow, 2010). Since the FFM is the most accepted model, it was the basis for the RPBF.

2.3 Use of the Big Five Model in the HRM context

The meta-analysis by Barrick and Mount (1991) was an important study in terms of the acceptance of the use of personality questionnaires as a predictor of job performance. Barrick and Mount (1991) used the FFM to study the relationship between the Big Five factors and a number of job-performance criteria for several occupational groups. In particular, they reported that Conscientiousness is a valid

predictor for most occupations, and that Extraversion is a predictor in positions in which interpersonal contact is important (e.g. sales positions). This was followed by a raft of studies on the relationship between personality and job success (Ones et al., 2007), which showed that other factors apart from Conscientiousness and Extraversion have a predictive value, for example: Agreeableness for customer service (Hurtz & Donovan, 2000), Openness to Experience for creativity and innovation (Grucza & Goldberg, 2007; Hough & Dilchert, 2007) and Emotional Stability for teamwork (Barrick, Mount, & Judge, 2001).

2.3.1 Facets

Apart from factor-level research, studies were also conducted into the relationship between the narrower personality dimensions (the 'facets') and job performance. Facets can be seen as more contextual manifestations of the broader factors (Roberts, 2006). Facets give a more complete and more detailed picture of someone's personality (Briggs, 1989; Mershon & Gorsuch, 1988). This is important in a work context because performance requirements apply in specific organizational contexts, such as hierarchical relationships (e.g., leadership), social situations (e.g., teamwork) or contributing to the company's results (e.g., sales targets). In many cases, the correlation between specific performance-related behaviour and one or more facets is stronger than the correlation obtained when factors are used. Facets can therefore show higher predictive validities (Ashton, 1998; Hough & Oswald, 2008; Paunonen, 1998). Research by Hough, Ones, & Viswesvaran (1998), Hough (1992) and Dudley, Orvis, Lebiecki, & Cortina (2006), for example, showed that the facets Dependability and Achievement have a stronger correlation than the factor Conscientiousness with criteria such as managerial performance, healthcare performance, and job dedication.

There is less consensus on the taxonomy of facets than there is on the five factors. The most well-known taxonomies are the 45-facet structure of the AB5C model (Hofstee, De Raad, & Goldberg, 1992) and the 30-facet structure of the NEO PI-R (Costa & McCrae, 1992), but other structures are also used (e.g., the TAPAS questionnaire with 22 facets; Stark, Drasgow, & Chernyshenko, 2008). Sometimes the linking of facets to the various factors is also inconsistent. In the NEO PI-R, the facet Warmth is placed under Extraversion, whereas in the AB5C model it is regarded as an aspect of Agreeableness.

2.3.2 Big Five in Different Contexts

Personality does not operate in a vacuum. People behave in different contexts such as at home, at work, at school, or together with friends at social gatherings. For a long time items in personality questionnaires were constructed context independent. However, contextualizing personality items in work settings has been found to enhance validity (e.g., Bing, Whanger, Davison, & VanHook, 2004; Truxillo, Bauer, Campion, & Paronto, 2002). The growing research on frame-of-reference effects in personality measurement suggests that the relationships between personality and work-related outcomes may be increased via the use of work-specific personality measures (Bing et al., 2004; Heller, Ferris, Brown, & Watson, 2009; Hunthausen, Truxillo, Bauer, & Hammer, 2003; Lievens, De Corte, & Schollaert, 2008). Unlike traditional personality measures that ask people to report how they behave in general, work-specific personality measures ask employees to report how they behave at work. In one study Heller et al. (2009) found that (1) work-specific personality measures predicted job satisfaction better than did general personality measures and (2) the effects of general personality descriptions on job satisfaction were mediated through work-specific personality. Other studies have found similar results using work-specific personality to predict job performance (Hunthausen et al., 2003) and school-specific personality to predict student performance (Bing et al., 2004; Lievens et al., 2008). The RBFP was also constructed as a work-related personality questionnaire.

2.4 Development of a Big Five Questionnaire for the Workplace

The RBFP is aimed at work-related behavior. That is not to say that it only refers to behavior in formal organizational contexts, but to all task contexts in which "performance goals" are important, be it paid or unpaid, work or leisure. In practice, however, the RBFP is predominantly used as an instrument supporting organizational HRM practices like personnel selection, training, or career development.

2.4.1 The Workplace Big Five Profile

The first version of the RBFP, the Workplace Big Five Profile (WB5P) was constructed by Howard and Howard (2001). They started with the NEO PI-R (Costa & McCrae, 1992). In their work as consultants with the NEO PI-R in organizational contexts Howard and Howard (2001, p. 5) realized that the NEO PI-R was not an ideal instrument because the language used was not related to a work context, there were many items that refer to the private domain, for example "When I am having my favorite foods, I tend to eat too much". Furthermore, the NEO PI-R was too long. To deal with these drawbacks the WB5P was constructed as a shorter work-related version of the NEO PI-R. The development of the items, the forming of subscales as well as the reliability, validity, and norm group information was extensively described in Howard and Howard (2001).

2.4.2 The Workplace Big Five 1.0

In 2002, the Company constructed a Dutch version of the WB5P, the Workplace Big Five 1.0 (WB1.0). This version was with respect to the items, facets, and factors, a translated replica of the WB5P. However, the WB1.0 was administered via the Internet, and the content and the structure of the report were adapted to Company users and in line with the style of the Company. The WB1.0 included a mapping from the resulting personality profile on a comprehensive set of competencies used by the Company in its consultancy practice, whereas the WB5P included a mapping on a different though partly comparable specific set of competencies used by Howard and Howard (2001). After collecting data to determine psychometric quantities and norms the Company started to use the WB1.0 in its own consultancy practice.

Both psychometric analyses and practical experiences with the questionnaire suggested a number of aspects that could be improved in a next version of the questionnaire. The facet and factor structure of the WB1.0 differed on some facets from that of the NEO PI-R, conforming to the WB5P structure as found by Howard and Howard (2001), but psychometric analyses suggested that a slightly different structure was found in the Dutch sample. Also, the number of items for some facets contained too few items so that the reliability was low. Furthermore, the used figures and numbers in the report, as well as the explaining texts in the report were not always interpreted correctly by the users.

To deal with these shortcomings a new version of the WB1.0 was developed, the Reflector Big Five Personality 2.0 (short: RBFP).

2.4.3 Reflector Big Five Personality 2.0

The RBFP as it is reported here is originally constructed as a Dutch version. As was discussed above, the structure of the RBFP is based on that of the NEO PI-R, but with items that refer to behavior "at work". A first set of such items was already available in the WB1.0 based upon the WB5P of Howard and Howard (2001). Furthermore, the Company used at the time of the development of the RBFP their own Big Five questionnaire which also contained a set of items referring to work situations, different from those in WB1.0. This instrument is called the Connector P. Data were collected on three instruments simultaneously with a same set of subjects: NEO PI-R, WB1.0, and the Connector P. Based upon some preliminary factor analyses, items were selected for an initial item pool for the RBFP with the following characteristics (1) the items reflect as much as possible the factor structure of the item set of the NEO PI-R data (2) each item should enable a reference to behavior "at work", and (3) each item should refer to observable behavior. Two subject experts independently selected from the total set of items used in the factor analyses an initial master set by visually inspecting the factor loadings. Linear regression of the commonly selected set of items on the NEO PI-R facet scores and cross checking the regression weights in two independent subsamples led to a provisional item pool for further use.

The original WB5P of Howard and Howard (2001) contained an Openness facet named "Scope"; this facet refers to attention to details versus a broader scope. Because in the Company's data this facet had high correlations with the Conscientiousness facets, it was removed from the RBFP. Furthermore, the facet Intellectual Autonomy was added to Openness. Although the NEO PI-R does not contain such a facet, it was reasoned that within a work and organizational context such a facet might be an empirically distinguishable one with possible predictive relevance for organizational criteria. Thus, based upon adjectives describing autonomous behavior and expert knowledge, a provisional set of items was phrased purporting to reflect the new facet and to be tested on convergence and discrimination within the total set.

The a priori placement of all facets did not exactly conform to the empirically factor analytic structure in the Company's data. Specifically, three of the 24 facets turned out to load on different factors. In fact, when computing a six factor solution they partly turned out to define a weak sixth factor. The items from the Need for stability facet Reticence also had substantial loadings on Extraversion and Openness. The items from the Agreeableness facets Agreement and Deference also loaded on Extraversion. A possible interpretation of these facets is "being in the background

avoiding taking charge". Although there were some cross loadings, the Company chose to keep these facets with their intended factors. This was done for practical reasons, because the former version of the RBFP as well as users of the NEO PI-R are used to this structure, and for practical applications it is desirable not to have too great discrepancies between the number of facets per factor.

The final version of the RBPF consists of 144 items, distributed over five scales (Need for Stability, Extraversion, Openness, Agreeableness, and Conscientiousness). In Table 2.1 these scales and their underlying facets are given. The items are scored on a 5-point Likert scale. The answer most indicative of the trait is scored '5' and the answer least indicative is scored '1'.

2.5 Research studies for the RBFP

Schakel, Smid, and Jaganjac (2007) reported various psychometric analyses. Using data from a representative sample of 1121 persons of the Dutch working population, Schakel et al. (2007) found $\alpha = .87$ for Need for Stability, $\alpha = .91$ for Extraversion, $\alpha = .90$ for Openness, $\alpha = .86$ for Agreeableness, and $\alpha = .93$ for Conscientiousness, based on the estimate for a linear combination (see Nunnally, 1978, p. 248). Test-retest reliability with a time interval between both test completions ranging from four weeks through more than a year equaled $r = .76$ for Need for Stability, $r = .78$ for Extraversion, $r = .74$ for Openness, $r = .70$ for Agreeableness, and $r = .70$ for Conscientiousness.

Furthermore, Schakel et al. (2007) investigated the correlational structure between the RBFP and the NEO-PI-R. Results showed that for Need for stability, Extraversion, and Conscientiousness the correlations are $r = .77$, $r = .78$, and $r = .81$, respectively. For Openness and Agreeableness the correlations were lower; $r = .34$ and $r = .48$, respectively. This was probably due to the different factorial structure for these factors for the RBFP compared to the NEO-PI-R. Agreeableness contains also some Extraversion facets in the RBFP and Openness in the NEO PI-R also contained the facet Aesthetics. To assess the validity of the RBFP, Egberink, Meijer, and Veldkamp (2010) applied a mixture version of the graded response model to investigate scalability and predictive validity for the Conscientiousness scale of the RBFP in a career development context. A four-class solution yielded the best interpretable results. The classes differed mainly with respect to their scores on the subscales Perfectionism and Concentration. Results showed that Conscientiousness may be qualitatively different for different groups of persons and

Table 2.1
Factor and facet structure of the RBFP.

Scale	Content
Need for stability	the extent to which we react emotionally to setbacks
N1 Sensitiveness	how much we worry about ourselves
N2 Intensity	how easily we get angry
N3 Interpretation	the extent to which we emphasize problems above solutions
N4 Rebound time	how much time we need to rebound from setbacks
N5 Reticence	the extent to which we feel uneasy in a group
Extraversion	the extent to which we actively maintain contact with others
E1 Enthusiasm	the extent to which we associate with others in a pleasant/personal way
E2 Sociability	how easily and how often we seek the company of others
E3 Energy mode	the degree of energy and the pace of work we show
E4 Taking charge	the extent to which we take the lead
E5 Directness	the extent to which we express our opinions directly
O Openness	the extent to which we look for new experiences and new ideas
O1 Imagination	the amount of new ideas and applications we come up with
O2 Complexity	the extent to which we approach matters in a complex/theoretical way
O3 Change	the amount of change we strive for
O4 Autonomy	the extent to which we show autonomy in our opinions and arguments
Accommodation	the extent to which we place other people's interests above our own
A1 Service	the extent to which we are interested in the needs/interests of others
A2 Agreement	the extent to which we try to avoid differences of opinion
A3 Deference	the extent to which we pursue personal recognition
A4 Trust of others	how easily we place our trust in others
A5 Tact	how carefully we choose our words
Conscientiousness	the extent to which we are organized and purposeful
C1 Perfectionism	the extent to which we strive for perfect results
C2 Organization	the extent to which we work in an organized and structured manner
C3 Drive	the extent to which we strive to achieve more and more
C4 Concentration	the extent to which our attention stays focused on a task
C5 Methodicalness	the extent to which we plan with foresight and in detail

that the predictive validity of the test scores improved for persons in different classes as compared to fitting a unidimensional IRT model. This type of validity research may be regarded as in line with suggestions provided by Borsboom, Mellenbergh, and van Heerden (2004). In their view, validation is testing the hypothesis that the theoretical attribute has a causal effect on test scores. One way to investigate this is to specify a psychometric model that describes different types of information, for example, information from introspection or restrospection protocols or from an analysis of the content of the items. Psychometric models for cognitive process were proposed by for example Embretson (1983) who related psychometric process models to test validation. Although Egberink et al. (2010) did not specify different processes; they tried to unravel the response process for different groups.

To be able to investigate the predictive validity of the RBFP, the Company developed their own competency model (Schakel et al., 2007). It consists of 43 competencies divided over 6 content domains. This competency model is the basis for the Company's 360 degree feedback instrument, called the Reflector 360. In this instrument each competency can be evaluated by means of five behavioral statements which can be filled out by different persons working with the person of interest (e.g., line manager, colleague, subordinate). Bartram (2005) conducted a meta-analysis based on 29 validation studies regarding the relationship between personality and competencies. Based on earlier research, he defined eight broad competency factors, referred to as the Great Eight. In a first attempt to investigate the predictive validity of the Company's competency model and to relate it to the literature, the 43 competencies were restructured into the Great Eight by content experts. In Table 2.2, the titles of the Great Eight, their hypothesized Big Five predictors based on Bartram (2005) and the different competencies from the Company's competency model are given.

Table 2.2
The Great Eight domains, their hypothesized Big Five predictors and related Company's competencies.

Competency domain	Hypothesized Big Five predictor	Competencies Company model
Leading and Deciding	Extraversion	Decisiveness, Initiative, Leadership, Coaching, Delegation, Group leadership
Supporting and Cooperating	Agreeableness	Sensitivity, Teamwork, Listening, Integrity
Interacting and Presenting	Extraversion	Sociability, Networking, Persuasiveness, Impact, Independence, Negotiating, Oral presentation, Oral communication
Analyzing and Interpreting	Openness	Written communication, Problem analysis, Judgement
Creating and Conceptualizing	Openness	Learning ability, Creativity, Vision
Organizing and Executing	Conscientiousness	Self-organization, Planning and organizing, Management control, Results orientation, Customer orientation, Quality orientation, Work standards, Discipline, Organizational loyalty
Adapting and Coping	Emotional Stability	Adaptability, Behavioral flexibility, Stress tolerance
Enterprising and Performing	Agreeableness (negative)	Tenacity, Self-development, Ambition, Entrepreneurship, Market orientation, Extra-organizational awareness, Organizational sensitivity

Based on data from 16862 persons, correlations were calculated between the Big Five factor scores and the competency scores. The Big Five factor scores were obtained from the RBFP filled out by the employee of interest. The competency scores were obtained from the evaluations made by the line manager using the Reflector 360. Since not every competency is evaluated for each employee of interest, the average score on the available competencies related to the Great Eight was taken. This is also the reason for different sample sizes for the different competency domains. Table 2.3 displays the correlations between the Big Five factor scores and the Great Eight competency scores.

Table 2.3
Correlations between the Big Five factor scores and the Great Eight competency scores.

Competency domain	*n*	NEE	EXT	OPE	AGR	CON
Leading and Deciding	503	-.18	.24	.26		
Supporting and Cooperating	1734	-.11	.12		.19	.13
Interacting and Presenting	587	-.27	.35	.32	.18	
Analyzing and Interpreting	444			.24		.19
Creating and Conceptualizing	3713	-.10	.10	.28		
Organizing and Executing	3734	-.09	.11	.19	.05	.27
Adapting and Coping	903	-.28		.11		
Enterprising and Performing	820	-.11	.13	.27	.12	.13

Note. *n* = sample size; NEE = Need for stability; EXT = Extraversion; OPE = Openness; AGR = Agreeableness; CON = Conscientiousness.

Only correlations equal or larger than .05 are displayed. Even though, the correlations are somewhat higher than the ones found by Bartram (2005), the hypothesized Big Five predictors of the Great Eight are similar. The only exception is the competency domain 'Enterprising and Performing', which has the highest correlation with Openness ($r = .27$) instead of Agreeableness ($r = .12$).

2.6 Online Administration, Processing, and Reporting

The RBPF is a computer-based Big Five personality questionnaire applied to situations and behavior in the workplace. The questionnaire is only available on the Internet via the Company's website. Registered individuals within an organization get a login code through which they can request the RBFP. When registered, the candidate receives an email with a unique hyperlink through which he or she can fill out the questionnaire. After completion, the answers are automatically scored and processed, and a report is generated. This report is sent electronically to the candidate.

This online administration requires a careful instruction towards the candidate as well as a clear format and content of the report so that the candidate can understand and interpret the test results in a correct way. To be able to do so, the

report is written in a non-technical language. Also, the Company guarantees, through certification of registered professionals that for every candidate a professional is available for a feedback session.

2.6.1 Use of the RBFP: Population and Context

As mentioned above, the RBFP is aimed at members of the 'normal' population in a workplace context and can be used in organizational HRM settings like personnel selection, training, or career development. Therefore the norm group is also selected from this population.

2.6.2 Qualification of Users

The qualification of the user of the RBFP depends on the context. A "user" is defined here as an individual who discusses the content and the results of the report with the candidate. A minimum requirement is that a user should be able to explain the test results to a candidate and the consequences of these results for the next steps in the (assessment) procedure. To be able to do this, besides knowledge of the context in which the questionnaire is used, the user should also have relevant knowledge with respect to the background, the content, and the meaning of the RBFP. Furthermore, the user should have interviewing skills for giving adequate feedback to the candidate. The Company demands certification of these knowledge and skills as a condition to administer the questionnaire. This certification is based on a successful completion of a certification training. Managers or HRM professionals working in a selection, training, or career development context are eligible for a certification training, irrespective of earlier academic qualifications.

2.7 Interpretation of Scale Scores

2.7.1 Item Responses

In the RBFP each facet is represented by a set of six items. It is important to note that each question is formulated in the third person singular. This intends to promote a mental set within the candidate to take the stance of an external observer (see, Hofstee, 1994). Furthermore, each item was constructed so that it referred as unambiguously as possible to a clearly observable set of behaviors, enabling the candidate to respond honestly whether or not he/she displays those behaviors. In

the online instruction the candidate is explicitly invited to answer honestly referring to his own interest.

2.7.2 Scores on Facets and Factors

The scores on both the facets and the factors are given in T-scores. T-scores are standard scores with a mean of 50 and a standard deviation of 10. Reference is hereby made to the most recent norm group. Within the report of the questionnaire itself the meaning and the interpretation of a T-score is given in non-technical straightforward language that can be understood by the user who has successfully completed a certification program. Figure 2.1 displays a part from the report that candidates receive.

2.7.3 Social Desirability

Morgeson et al. (2007a) point out that, when personality questionnaires are used in situations in which the resulting scores have important consequences for the respondent, he/she might tend to give socially desirable answers to the questions. MacCann, Ziegler and Roberts (2011) refer to the use of "warnings" as a usable method for reducing faking. Therefore, the RBPF warns respondents that it is in their own interest to answer the questions as fully and truthfully as possible, in order to obtain an accurate picture of their potential for success at work.

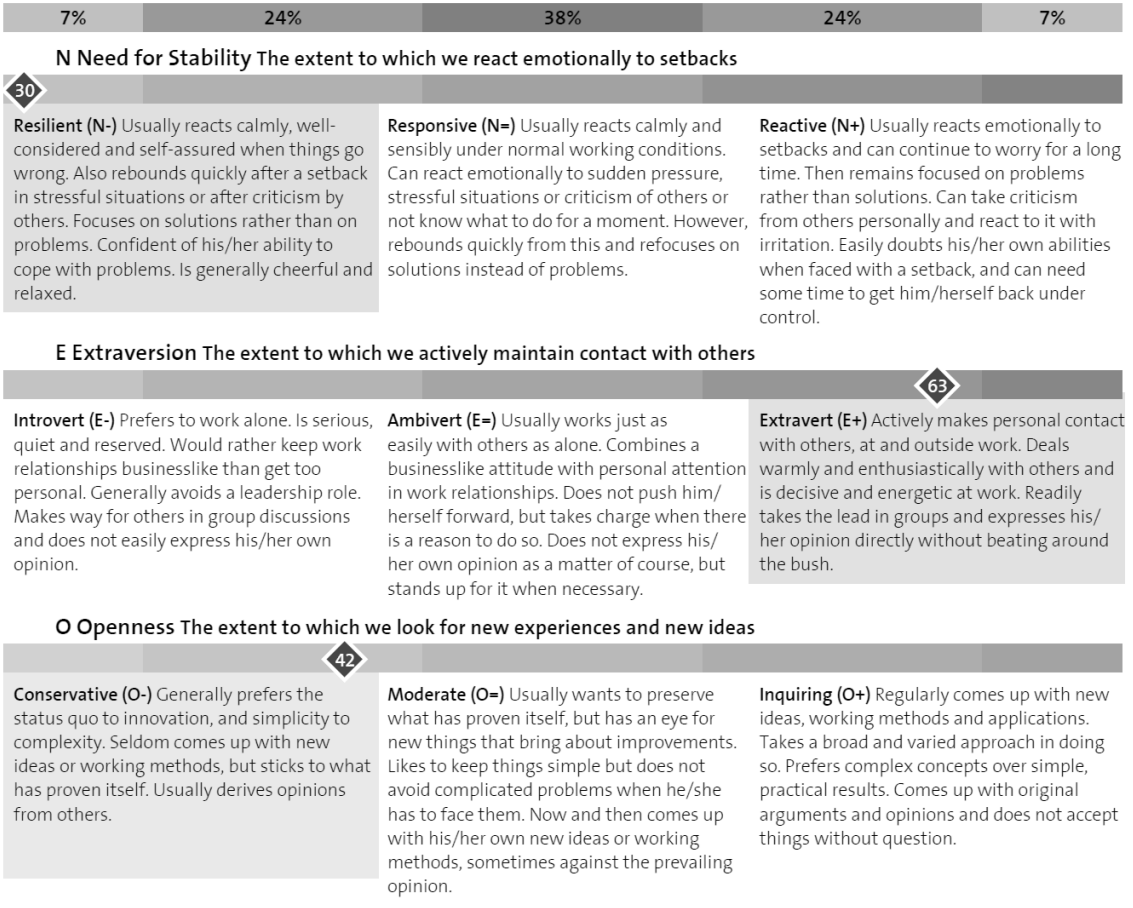


Figure 2.1: A part from the RBFP report.

Chapter 3
The Use of Effect Size Indices for Differential Item and Test Functioning in a Business Context

3.1 Introduction

Many employment firms and government agencies use the same questionnaire for different purposes, for example, for personnel selection and for employees training and development. Also, with the increasing use of online testing the administration of the same questionnaire to different (ethnic) groups is becoming the rule more than the exception. In these situations it is important to determine whether items and scales function similarly when administered for different test purposes or different groups, that is, it is important to determine measurement invariance.

Both confirmatory factor analytic methods and item response theory (IRT; Embretson & Reise, 2000) methods have been proposed to investigate measurement invariance. In this study we focus on IRT-based methods. One way to investigate whether the psychometric quality of a scale is comparable across different contexts is to apply IRT-based differential item functioning (DIF) and differential test functioning (DTF) techniques. However, most differential functioning (DF) studies only report statistical significant test results without reporting any effect size measures. Because DF results may be statistical significant without having much practical implications, recently, several authors have argued that when conducting DF research some kind of effect size should be reported. Stark, Chernyshenko, and Drasgow (2004) proposed a number of DF effect size measures and concluded in their research "that although many items exhibited bias in analyses of the large samples, the net magnitudes of effect on potential selection decisions were

This chapter has been submitted for publication.

nugatory" (p. 497). Meade (2010) discussed different types of DF effect size measures and proposed a taxonomy for group mean comparisons and for the comparison of individual respondents across different groups. Meade (2010) concluded his study by noting that "over the past two decades, significant progress has been made with methods of detecting statistically significant DF. However, a broader understanding and utilization of DF effect size is an essential next step in the progression of understanding invariance" (p. 740).

The aim of the current chapter is twofold. First, we evaluate whether the Reflector Big Five Personality (RBFP) shows measurement invariance properties across different populations. Second, we evaluate and compare the usefulness of different types of effect size measures. Thus, we analyze data from a personality questionnaire administered in a selection and a career development context and we report and compare a large number of different effect size indices. As such we contribute to the further understanding of differential item and test functioning in personnel selection and assessment.

This chapter is organized as follows. We first discuss a likelihood ratio DIF approach and different types of effect size indices. Second, we report DF analyses for the RBFP administered in different administration contexts and for different ethnic groups and, finally, we discuss the practical consequences for personality assessment, especially in the light of new developments in test construction, such as the construction of short scales.

Our study is framed in an IRT context, and because IRT is rapidly becoming the standard analysis technique nowadays for test and questionnaire construction, we will not explain the basic principles of IRT. The interested reader is referred to Embretson and Reise (2000) or Sijtsma and Molenaar (2002) for excellent introductions to parametric and nonparametric IRT approaches.

3.2 Differential Functioning of Items and Scales

3.2.1 Likelihood Ratio Approach

Many different IRT-based approaches have been proposed to investigate DF (for an overview see Millsap & Everson, 1993; Raju, van der Linden, & Fleer, 1995). In this study, we investigate DF using the popular likelihood ratio test (LRT) proposed by Thissen, Steinberg, and Wainer (1988, 1993). The LRT compares the fit (i.e., the likelihood) of a compact model in which all item parameters of all items are assumed

to be equal across both groups with the fit of an augmented model in which all item parameters of all items are assumed to be equal for both groups except for one item at a time. An overall significance test of DIF can be conducted by the test statistic:

G^2 (df) = -2 ($\ln L_C$ - $\ln L_A$), where

df is the degrees of freedom (i.e., number of item parameters), $\ln L_C$ is the log-likelihood of the compact model, and $\ln L_A$ is the log-likelihood of the augmented model. An item exhibits significant DIF when G^2 exceeds a critical value of the χ^2 distribution at a prespecified level of Type I error. This means that the augmented model fits better than the compact model for that item, suggesting that it is better to use different item parameters for both groups.

Although the LR method has been applied successfully (Bolt, 2002; Meade & Lautenschlager, 2004), its power is high with large sample sizes. Then, statistical testing is not sufficient to determine practical importance, because small differences may not be very relevant in practice. Therefore, different effect size measures have been proposed that can be used to judge whether significant differences have practical meaning. For example, Stark, Chernyshenko and Drasgow (2004) discussed two different methods: an effect size measure for the raw score and an effect size measure using the ratio of selection ratios. By means of these methods they investigated the impact of DIF and DTF on potential selection decisions when comparing the scores of applicants and nonapplicants on personality scales. Although their results showed that a lot of items exhibit DIF, the overall effect on selection decisions was small. Recently, Meade (2010) presented a taxonomy of different effect size measures for differential item and test functioning. In the present study we apply several of these indices.

3.2.2 Description of Effect Size Indices

Meade (2010) used four criteria on the basis of which different effect size indices were distinguished: (1) DF on the item and/or scale level (2) DF cancels across items and/or latent trait values (3) DF are reported in the original metric or normed to a standard deviation metric, and (4) DF on the basis of a sample distribution or on the basis of an assumed theoretical distribution.

In the present study we focus on polytomous item scores and we use the same notation as in Meade (2010). All the indices use the expected score (ES) for respondent s (s =1, …, N), with an estimated latent trait value, $\widehat{\theta}$, for item i (i =1, …, j). This ES equals the sum of the probabilities of a response to each of the k = 1, …, m response options times the value of that response option X_{ik}, that is,

$$ES_{s(\hat{\theta})i} = \sum_{k=1}^{m} P_{ik}(\hat{\theta})\, X_{ik} \ .$$

The expected score is similar to an item level true score and has a potential range from the lowest response option to the highest response option. Similarly, the expected *test* score (ETS) equals:

$$ETS_s = \sum_{i=1}^{j} ES_{si} \ .$$

The indices can be used to investigate whether the items and test function differently in a focal (often a minority group, for example, Blacks) and a reference group (majority group, for example, Whites). To be able to do this, first item parameters are estimated in both groups separately and linked to a common metric via, for example, concurrent calibration as is done in the LRT approach. Once linked, each item is associated with two sets of item parameters, one set associated with the focal group and one set associated with the reference group. In general, the minority (or the group with the lowest score) is chosen as the focal group and the majority (or the group with the highest score) as the reference group (e.g., Stark et al., 2004). In the present study we chose the incumbents as the focal group and the applicants as the reference group. After the parameters were estimated for both groups, the ESs were compared for the focal and reference group.

A simple effect size index at the item level is the average difference in ESs across the persons in the focal group sample. This index, the signed item difference in the sample ($SIDS_i$), equals:

$$SIDS_i = \frac{\sum_{S=1}^{N}\left[ES_{(si|\hat{\theta},\gamma F)} - ES_{(si|\hat{\theta},\gamma R)}\right]}{N} \quad \text{with} \qquad (1)$$

γF = the estimated item parameters in the focal group, and

γR = the estimated item parameters in the reference group.

The sum of these differences across the *j* items will result in a scale level index: the signed test difference in the sample (STDS):

$$STDS = \sum_{i=1}^{j} SIDS_i \ . \qquad (2)$$

Both indices use the sample distribution and display the differences in the original metric. This means that when, for example, for a five category item SIDS = -2.2, it is expected that persons in the focal group will score 2.2 points lower on that item than persons in the reference group with the same latent trait value. For the STDS this difference is related to the difference in summed scale score. The SIDS allows for cancellation of DF across persons and the STDS allows for cancellation across items and persons. At the item level this implies that the SIDS might indicate that there is no DF present, whereas DF might be present at different trait levels, but

that the sum of these differences equals zero. At the scale level, this form of cancellation can also take place across items.

To prevent cancellation across items and, or persons, absolute bars can be used in Equations 1 and 2, resulting in the unsigned item difference in the sample (UIDS) and the unsigned test difference in the sample (UTDS):

$UIDS_i = \frac{\sum_{s=1}^{N} \left| ES_{(si|\hat{\theta},\gamma F)} - ES_{(si|\hat{\theta},\gamma R)} \right|}{N}$, and

$UTDS = \sum_{i=1}^{j} UIDS_i$.

Like the SIDS and the STDS, the UIDS and the UTDS use the sample distribution and display the differences in the original metric. The difference is that the UIDS does not allow cancellation across persons and the UTDS does not allow cancellation across items and persons. UIDS can be interpreted as the hypothetical difference in ESs had the DF in that item been uniform across persons, which means always favoring one group. UTDS can be interpreted in the same way, but now at the test level.

The indices described above all report the differences in the original metric. A standardized difference at the item level can be computed by the expected score standardized difference (ESSD) and this difference can be reported at the test level by the expected test score standardized difference (ETSSD):

$ESSD_i = \frac{\overline{ES}_{(\gamma F)} - \overline{ES}_{(\gamma R)}}{SD_{ItemPooled}}$ with

$\overline{ES}_{(\gamma F)}$ = mean ES using the estimated item parameters in the focal group, and

$\overline{ES}_{(\gamma R)}$ = mean ES using the estimated item parameters in the reference group, and

$ETSSD = \frac{\overline{ETS}_{(\gamma F)} - \overline{ETS}_{(\gamma R)}}{SD_{TestPooled}}$ with

$\overline{ETS}_{(\gamma F)}$ = mean ETS using the estimated item parameters in the focal group, and

$\overline{ETS}_{(\gamma R)}$ = mean ETS using the estimated item parameters in the reference group.

The differences are normed to a standard deviation metric and can, therefore, be interpreted using Cohen's (1988) rules of thumb for small, medium, and large effect sizes (Meade, 2010). These indices also use the sample distribution. Like SIDS and STDS, ESSD allows for cancellation across persons and ETSSD allows for cancellation across both items and persons.

Finally, we discuss the unsigned expected test score difference in the sample (UETSDS):

$$UETSDS = \frac{\sum_{S=1}^{N}\left[\left|ETS_{(s|\hat{\theta},\gamma F)} - ETS_{(s|\hat{\theta},\gamma R)}\right|\right]}{N}.$$

This index at the scale level differs from the other scale level indices, because it allows cancellation across items (because ETS is the sum of the item ESs), but not across persons (due to the absolute bars). Like the other indices, the sample distribution is used instead of an assumed theoretical distribution. The differences in ETSs are displayed in the metric of observed scores. UETSDS can be interpreted as the hypothetical amount of DF at the scale level had the DF been unidirectional in nature, which means always favoring one group.

Meade (2010) suggested that researchers should always report the STDS, UETSDS, and the ETSSD regardless of their research purposes. Comparing STDS and UETSDS provides information with regard to cancellation of DF across the trait score. When STDS and UETSDS are equal, cancellation of DF might occur across items, but it does not occur across the latent trait. The ETSSD is very useful since the differences in ETSs are normed to a standardized metric and this index can be used for tests containing items with different numbers of response categories. Besides examining the effect size indices, we visually inspected the ES and ETS plots.

In this study we only used indices based on the sample distribution, for the use of indices based on an assumed theoretical distribution, we refer to Appendix B in Meade (2010).

3.3 Method

3.3.1 Instruments

Reflector Big Five Personality (RBFP)

As discussed in Chapter 2 the RBFP (Schakel, Smid, & Jaganjac, 2007) is a computer-based Big Five personality questionnaire applied to situations and behavior in the workplace. It consists of 144 items, distributed over five scales (Need for stability, Extraversion, Openness, Agreeableness, and Conscientiousness). The items are scored on a five point Likert scale. The answer most indicative for the trait being measured is scored '4' and the answer least indicative for the trait is scored '0'. For this study, we selected the Need for stability scale, the Extraversion scale, and the Conscientiousness scale. For this study, we recoded the Need for stability scale such that it can be interpreted as an Emotional Stability scale. In that

way, all three scales are phrased in a positive direction. Each scale consists of 30 items, equally distributed over five subscales. The RBFP is a Dutch version of the Workplace Big Five Profile constructed by Howard and Howard (2001). This profile is based on the NEO-PI-R (Costa & McCrae, 1992) and adapted to workplace situations. For the Dutch version, both conceptual analyses and exploratory factor analyses showed the Big Five structure (Schakel et al., 2007).

3.3.2 Sample and Procedure

Data were collected between September 2009 and January 2011 by the Company. We distinguish two groups: (1) applicants who apply for a job at an organization, and (2) incumbents who already work for an organization and completed the RBFP as part of their own personal career development. We used data from 4050 applicants (M_{age} = 33.5, SD = 9.23); 62.1% men; 80.9% native, 9.4% Western immigrants and 9.6% non-Western immigrants. 34.6% of the participants had a university degree, 44.7% had higher education, and 20.7% secondary education. Data from the incumbents consisted of 4217 persons (M_{age} = 39.4, SD = 9.31); 55.0% men; 88.8% native, 7.0% Western immigrants and 4.2% non-Western immigrants. 27.6% of the participants had a university degree, 49.4% had higher education, and 23.0% secondary education.

3.3.3 Analysis

Item and Scale Quality

To obtain an impression of the item and scale quality, we used the computer program Mokken Scale Analysis for Polytomous items (MSP5.0; Molenaar & Sijtsma, 2000). MSP5.0 is a handy tool to obtain a first impression about the quality of the data. The program contains several easy-to-interpret statistics like item proportion correct score reflecting item difficulty, and scalability coefficients (H) reflecting discrimination power. Besides at the scale level, H is also defined at the level of the item(step)-pair level (H_{ij}) and item level (H_i), and can be expressed in terms of observed versus expected number of Guttman errors or in terms of observed versus maximal possible covariance between items (for exact formulas, see e.g. Sijtsma & Molenaar, 2002, pp. 51-58).

For the interpretation of H, Sijtsma and Molenaar (2002, pp. 60) give the following guidelines. The scale H should be above .3 for the items to form a scale. When $.3 \leq H < .4$ the scale is considered weak, when $.4 \leq H < .5$ the scale is

considered medium, and when $H \geq .5$ the scale is considered strong. In addition, although point estimates of θ cannot be obtained, an estimated ordering of subjects by their estimated θ-values is possible using the number-correct score. Examples of applications of Mokken scaling in the typical performance domain can be found in, for example, Meijer and Baneke (2004) who showed the usefulness of Mokken scaling to analyze the MMPI depression scale, Moorer, Suurmeijer, Foets, and Molenaar (2001) who applied Mokken scaling to the Rand-36, and Meijer, Egberink, Emons, and Sijtsma (2008) who discussed the use of Mokken scaling to identify atypical response behavior.

Besides comparing the scalability of the items for applicants and incumbents, we also used MSP5.0 to investigate differential item functioning. MSP5.0 has incorporated a simple procedure to provide plots for a visual inspection of DF using the splitter item method. Using this method the sample can be split on the basis of, for example, item values of a group variable, like boys and girls, and it permits a fast overview of the item means when there are two subgroups. Although the plots provided by MSP5.0 are informative with respect to the ordering of the items, these plots do not provide information conditional on the latent trait value. By means of the LRT, DF can be investigated conditional on the estimated latent trait score.

Differential Item and Test Functioning

DF was investigated across two groups within different administration contexts: applicants within a selection context (reference group) and incumbents within a career development context (focal group). We also investigated DF for different ethnicity groups[1] in a selection context. We compared the following groups: 1. Dutch natives (reference group) and Western immigrants (focal group), 2. Dutch natives (reference group) and non-Western immigrants (focal group), 3. Western immigrants (reference group) and non-Western immigrants (focal group).

We used the program IRTLRDIF (Thissen, 2001) to determine statistical significant DF, to estimate the item parameters for both groups, and to link them to a common metric. In IRTLRDIF the three-parameter logistic model and the graded

[1] According to the Dutch Central Bureau of Statistics (Centraal Bureau voor de Statistiek, 2000), Dutch natives are citizens who are born in the Netherlands, just like their parents. Western immigrants are born in western, northern or southern Europe, the USA, Canada, Australia, New Zealand, Japan or Israel and non-Western immigrants are born in one of all other countries. First-generation immigrants are born abroad, just like at least one of the parents. Second-generation immigrants are born in the Netherlands, but at least one of their parents is born abroad. In this study, we did not distinguish between first- and second-generation immigrants, since the different groups would be too small for the analyses.

response model are implemented. For this study we used the graded response model (Samejima, 1969, 1997). This model has been often applied to personality data (e.g., Ankenmann, Witt, & Dunbar, 1999; Embretson & Reise, 2000). The estimated and linked parameters together with the focal group data were then used as input for VisualDF (Meade, 2010) which can be downloaded from http://www4.ncsu.edu/~awmeade.

For the comparison of the mean scores of the applicants and the incumbents, cancellation of DF across items and persons is appropriate. Therefore, we used SIDS and ESSD at the item level, and STDS and ETSSD at the scale level. For the comparison of different ethnic groups in a selection context, cancellation across items is appropriate, but not across the latent trait. Therefore, we used UIDS at the item level, and UTDS and UETSDS at the scale level. Furthermore, we compared SIDS and STDS to UIDS and UTDS to assess the extent to which cancellation of DF across items and trait values occurs.

3.4 Results

3.4.1 Descriptive Statistics

Table 3.1 shows descriptive statistics for the three scales Emotional Stability, Extraversion, and Conscientiousness. From this table it is clear that, in general, there were no large differences in the scalability of the items for the applicants and the incumbents. However, for the Emotional Stability scale the overall H value was .03 lower for the applicants than for the incumbents. Although the reliability of the scales is high (α between .86 and .90), there were also some items in each scale with relatively low H_i values (smaller than H_i =.30), indicating that these items did not relate strongly to the underlying latent trait.

Table 3.1 also shows that the mean scores for the applicants were higher than for the incumbents, especially for Emotional Stability and Conscientiousness, resulting in medium effect sizes. To further investigate the effect of different mean scores, we plotted the item means for the applicants against the item means for the incumbents (Figures 3.1a through 3.1c). MSP5.0 provides these plots using the splitter item method (Molenaar & Sijtsma, 2000). We observe that for the Extraversion scale (Figure 3.1b) most items are on or near the diagonal with few outliers (which would indicate that an item is much more popular in one group), indicating that there is no relation with the group variable administration context

Table 3.1

Descriptive statistics for the Emotional Stability, Extraversion, and Conscientiousness scales.

	applicants						incumbents						
Scale	α	range item-test correlations	*M*	*SD*	*H*	range H_i values	α	range item-test correlations	*M*	*SD*	*H*	range H_i values	*d*
EMS	.88	.12 - .61	87.87	12.13	.24	.06 - .33	.90	.20 - .63	81.43	14.70	.27	.12 - .35	.48
EXT	.86	.15 - .56	85.55	11.17	.20	.08 - .28	.86	.20 - .57	82.66	12.93	.20	.11 - .29	.24
CON	.90	.27 - .59	93.36	12.07	.27	.16 - .34	.90	.24 - .60	86.21	14.11	.26	.14 - .34	.54

Note. EMS = Emotional Stability; EXT = Extraversion; CON = Conscientiousness; *d* = Cohen's *d* for difference in mean scale score between both groups.

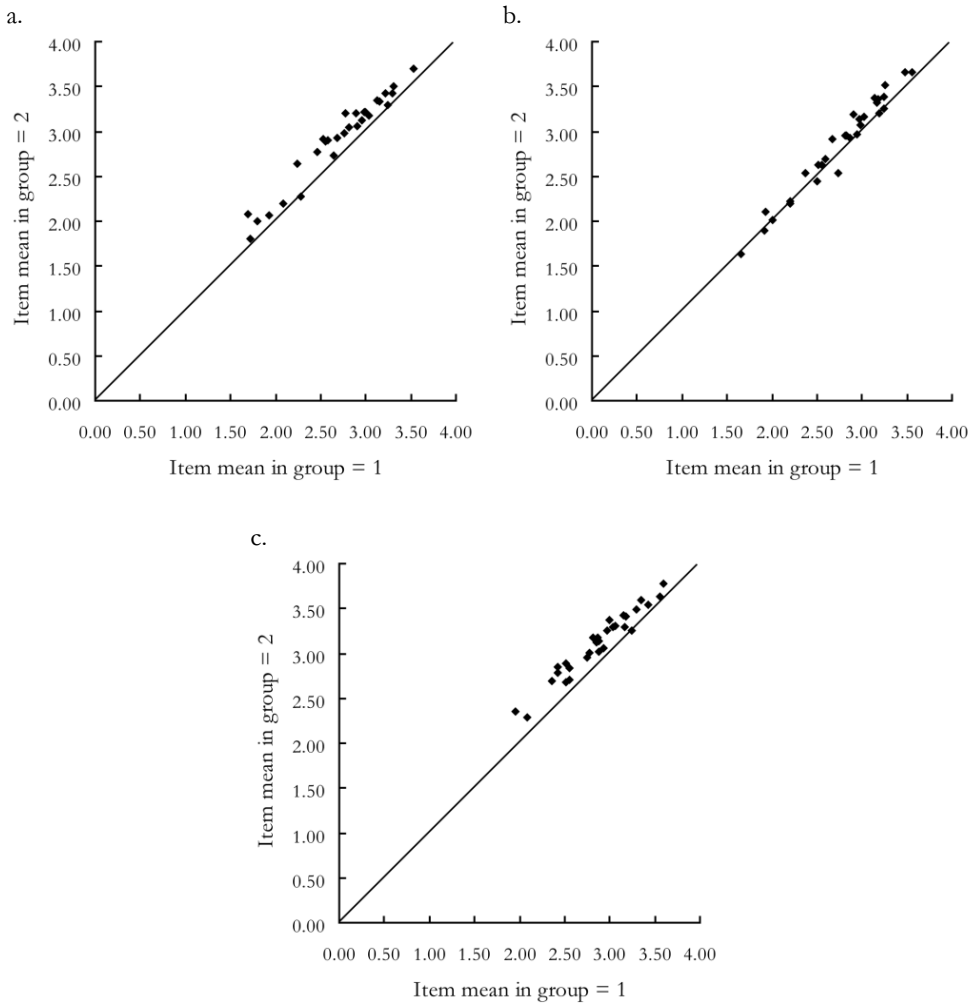


Figure 3.1: Mean splitter plot for group 1 (x-axis, incumbents) vs group 2 (y-axis, applicants) for the Emotional Stability items (a), the Extraversion items (b), and the Conscientiousness items (c).

(i.e., selection/career development). Note, however, that the item means for the Emotional Stability scale and Conscientiousness scale (Figures 3.1a and 3.1c) are almost all above the diagonal line indicating that many items are more popular for applicants in the selection group than for incumbents in the developmental group.

Table 3.2

H and H_i values for the selected Emotional Stability, Extraversion, and Conscientiousness items.

	EMS		EXT		CON	
	applicants	incumbents	applicants	incumbents	applicants	incumbents
item 1	.32	.34			.36	.34
item 2			.29	.29	.27	.27
item 3	.28	.30			.29	.29
item 4	.35	.36	.30	.26	.21	.22
item 5	.32	.35			.36	.36
item 6	.23	.30			.34	.32
item 7	.20	.22	.34	.33	.25	.28
item 8	.21	.25	.20	.19	.26	.29
item 9	.26	.27	.32	.33	.31	.31
item 10	.33	.31	.27	.23	.35	.34
item 11	.28	.30			.36	.36
item 12	.30	.32	.24	.23	.36	.35
item 13	.27	.29	.31	.28	.28	.27
item 14	.27	.28				
item 15	.36	.35	.28	.27		
item 16	.31	.33	.22	.21		
item 17	.25	.29				
item 18	.26	.27			.30	.28
item 19	.23	.28	.26	.26	.23	.27
item 20	.32	.35	.34	.36	.29	.29
item 21	.36	.35	.26	.28	.28	.29
item 22			.29	.32	.31	.32
item 23	.30	.28			.32	.31
item 24	.35	.37	.34	.33	.21	.25
item 25			.23	.25	.21	.20
item 26	.23	.23	.27	.32	.33	.31
item 27	.33	.32	.26	.28	.21	.20
item 28			.28	.31	.35	.33
item 29	.19	.21	.22	.26	.31	.28
item 30	.27	.27	.27	.31	.34	.34
H	.28	.30	.28	.28	.30	.30

Note. EMS = Emotional Stability; EXT = Extraversion; CON = Conscientiousness.

Although, the scalability of the scales is comparable for both groups, the relatively low H_i values for some items lead to item estimation problems for the DF analyses. Therefore, we ran the SEARCH option in MSP5.0 with different lower bounds (i.e., c = .30, c = .25 and c = .20) to explore item quality. On the basis of these results, items with H < .20 were removed from the scale. Table 3.2 depicts the H_i values for the selected items of each scale. These items will be used for the DF analyses.

3.4.2 Differential Functioning (DF) analyses

Comparing group means

The shift in item means for the applicants may be due to a shift in the latent trait distribution, but items may also show DF in the two groups. To investigate DF, we calculated the likelihood ratio statistic and several effect size measures. Although many items showed statistically significant DIF according to the LRT at the 5% level, we inspected the values on the different effect size indices to obtain an impression about the practical importance of these significant results. The values of the effect size indices are given in Table 3.3. Because cancellation of DF is appropriate across trait values when comparing group means, it is recommended to use SIDS and ESSD at the item level, and STDS and ETSSD at the test level.

The results depicted in Table 3.3 show that items from each scale have positive and negative SIDS values, which indicates that for some items the applicants are in favor (i.e., negative SIDS value) and for some items the incumbents are in favor (i.e., positive SIDS value). For the Emotional Stability scale, the highest negative SIDS value is -0.14 (items 3 and 11) and the highest positive SIDS value is 0.10 (items 4 and 29). This means that for items 3 and 11 the incumbents scored 0.14 points lower than the applicants, and for items 4 and 29 the incumbents scored 0.10 points higher than the applicants. The highest negative and positive SIDS values are -0.15 and .11 for the Extraversion scale, and -.15 and .10 for the Conscientiousness scale. Note that SIDS values are reported in the item expected score metric and that the items have five response categories, scored 0-4, which suggests that the found differences are small. This is confirmed by the ESSD values, which are standardized. The results show that all significant differences identified by the LRT are small (i.e., $|ESSD| < 0.30$), with the exception of items 10, 13, 15, and 25 from the Extraversion scale. These differences are of medium effect size (i.e., $0.30 < |ESSD| < 0.70$).

Table 3.3

Item- and scale-level effect size statistics for the selected items of the Emotional Stability, Extraversion and Conscientiousness scales for the incumbents-applicants comparison.

	Emotional Stability		Extraversion		Conscientiousness	
item	SIDS	ESSD	SIDS	ESSD	SIDS	ESSD
1	-0.13	-0.23			0.02	0.03
2			-0.03	-0.08	0.10	0.20
3	-0.14	-0.28			-0.06	-0.11
4	0.10	0.15	-0.06	-0.18	0.01	0.03
5	0.05	0.13			-0.04	-0.06
6	0.06	0.13			0.07	0.12
7	-0.08	-0.25	-0.09	-0.19	-0.07	-0.16
8	0.04	0.09	-0.03	-0.15	-0.03	-0.08
9	-0.10	-0.22	-0.01	-0.01	0.05	0.09
10	0.02	0.04	-0.15	-0.34	-0.10	-0.18
11	-0.14	-0.25			0.02	0.03
12	-0.06	-0.11	0.08	0.25	-0.01	-0.02
13	-0.03	-0.07	-0.15	-0.42	-0.06	-0.25
14	-0.06	-0.20				
15	-0.10	-0.16	-0.11	-0.39		
16	0.08	0.14	-0.03	-0.07		
17	0.02	0.04				
18	-0.02	-0.05			-0.03	-0.07
19	-0.01	-0.03	0.00	-0.01	-0.09	-0.21
20	-0.04	-0.07	-0.04	-0.05	0.08	0.15
21	0.00	-0.01	-0.06	-0.12	-0.13	-0.25
22			0.08	0.21	-0.15	-0.26
23	0.07	0.15			-0.09	-0.16
24	0.02	0.04	0.08	0.14	0.05	0.13
25			0.11	0.30	0.04	0.14
26	-0.11	-0.23	0.07	0.14	0.05	0.11
27	0.04	0.08	0.10	0.24	0.00	0.01
28			0.11	0.25	0.00	0.00
29	0.10	0.27	0.01	0.03	0.05	0.15
30	-0.01	-0.03	0.07	0.18	0.03	0.06
STDS	-0.45		-0.03		-0.28	
UETSDS	0.47		0.12		0.31	
ETSSD	-0.04		0.00		-0.02	

Note. SIDS = signed item difference in sample; ESSD = expected score standardized difference; STDS = signed test difference in the sample; UETSDS = unsigned expected test score difference in sample; ETSSD = expected test score standardized difference.

At the test level the STDS values indicate that the incumbents scored 0.45 points lower than the applicants on the Emotional Stability scale, 0.03 points lower on the Extraversion scale, and 0.28 points lower on the Conscientiousness scale. Note that the range of scores in this sample for the Emotional Stability scale is 15-104, for the Extraversion scale is 16-84, and for the Conscientiousness scale is 17-104, which suggests that the observed differences are small. This is also confirmed by the values of the standardized effect size, ETSSD which are -0.04, 0.00, and -0.02 for the Emotional Stability scale, the Extraversion scale, and Conscientiousness scale, respectively.

Furthermore, all scales have positive and negative SIDS values, which indicate cancellation of DF across items. However, since STDS and UETSDS values are comparable, there is no cancellation across persons at the scale level. The difference between those values is largest for the Extraversion scale (i.e., STDS = -0.03 and UETSDS = 0.12). However, after inspecting the ETS curves which were almost identical, we conclude that cancellation of DF across persons is negligible.

Comparing different ethnic groups

Table 3.4 summarizes the results with respect to the DF analyses for different ethnic groups. For the selection context, we compared Dutch natives, Western immigrants, and non-Western immigrants. The results of the LRT showed, again, many statistically significant DF results (third column in Table 3.4). Because cancellation across persons is not appropriate in a selection context, UIDS is used at the item level and UTDS and UETSDS at the scale level. Also UIDS and SIDS, and UTDS and STDS are compared to assess whether there is cancellation across persons.

Comparison of the SIDS and UIDS indices showed that there was cancellation of DF across persons for some items in each comparison for the three scales (fourth column in Table 3.4). Figure 3.2 shows the ES plots for two items of the Emotional Stability scale. The plot of Item 8 'Is even tempered' (upper panel) shows that for the lower latent trait values (i.e., $\theta < -1.75$) non-Western immigrants are in favor in comparison with natives and that the opposite is true for the higher trait values (i.e., $\theta > -1.75$). Thus, there was cancellation of DF across persons for this item. However, one may argue that this does not really represent much of a reversal because θ values less than -1.75 will be rare, so there will be few persons for whom the Non-Western immigrants has a higher ES. The plot of Item 29 'Does not hesitate to express his/her opinion' (lower panel) shows that there is no cancellation of DF across persons, because the functions do not intersect. Natives are always in

Table 3.4
Results of the DF analyses for the different ethnicity groups in a selection context.

			VisualDF				
scale	comparison	LRT	SIDS/ UIDS	STDS	UTDS	UETSDS	ETSSD
Emotional Stability	natives vs non-Western immigrants	25 items	7 items	-0.54	2.88	0.54	-0.05
26 items	natives vs Western immigrants	22 items	9 items	-0.29	1.52	0.29	-0.03
	Western vs non-Western immigrants	21 items	11 items	-0.37	2.31	0.38	-0.04
Extraversion	natives vs non-Western immigrants	20 items	9 items	0.13	2.31	0.18	0.02
21 items	natives vs Western immigrants	15 items	12 items	-0.08	0.96	0.08	-0.01
	Western vs non-Western immigrants	15 items	10 items	0.20	2.17	0.20	0.03
Conscientiousness	natives vs non-Western immigrants	23 items	5 items	-0.05	2.41	0.17	-0.01
26 items	natives vs Western immigrants	19 items	14 items	0.02	1.22	0.07	0.00
	Western vs non-Western immigrants	21 items	7 items	-0.07	2.13	0.12	-0.01

Note. LRT = likelihood ratio test, number of items flagged as significant DIF at the 5% level; SIDS/UIDS = number of items with different values for the signed item difference in sample and the unsigned item difference in sample that indicates cancellation across theta; UTDS = unsigned test difference in the sample; UETSDS = unsigned expected test score difference in sample.

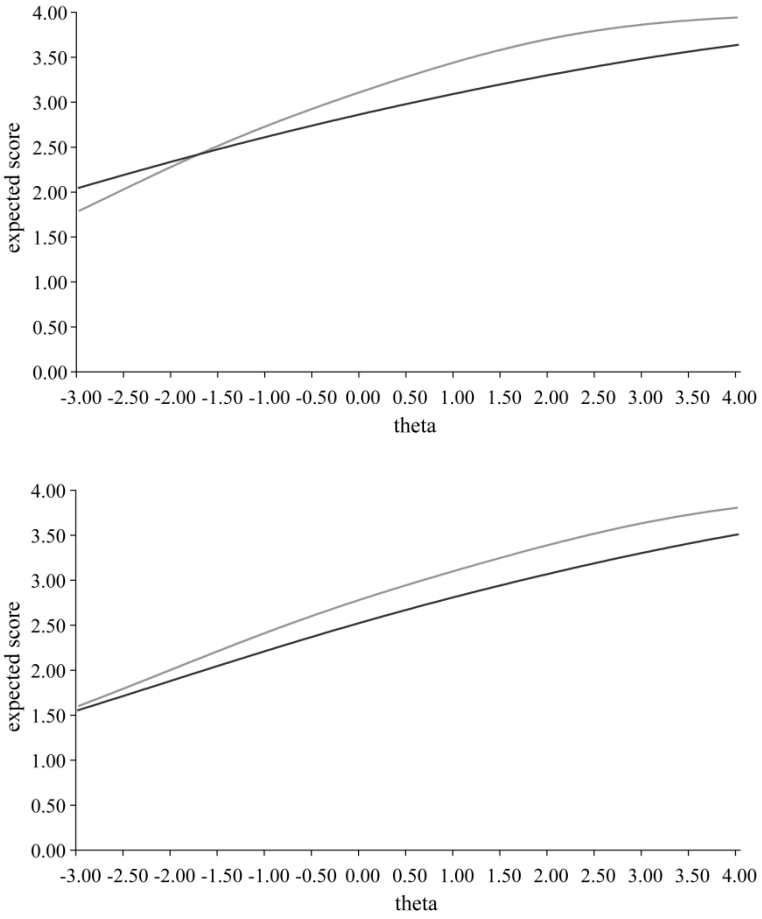


Figure 3.2: Expected score plots of Item 8 'Is even tempered' (upper panel; SIDS = -0.26, UIDS = 0.26, ESSD = - 0.90) and Item 29 'Does not hesitate to express his/her opinion' (lower panel; SIDS = -0.26, UIDS = 0.26, ESSD = -0.84) for the Emotional Stability scale for the comparison of natives and non-Western immigrants. *Note.* The black line represents the item means for the Non-Western immigrants (=focal group) and the grey line represents the item means for the natives (=reference group).

favor in comparison with the non-Western immigrants on this item, that is, natives always score higher on this item.

At the scale level, the UTDS indicates the hypothetical difference in ETSs had the DF been uniform across persons, which means always favoring one group. For example, for the comparison between the Western and the non-Western immigrants on the Extraversion scale the UTDS is 2.17, which means that had the DF been

uniform one of the two groups would have been expected to score on average 2.17 points higher than the other group. The UTDS and STDS values are not equal, indicating that cancellation across items and/or respondents occurred. To assess whether cancellation occurred across items and persons, or across items and not across respondents, or across respondents and not across items, comparing SIDS with UIDS values, and STDS with UETSDS provides more information. The absolute values of the SIDS are not equal to the corresponding UIDS for many items (see fourth column of Table 3.4), suggesting that cancellation across respondents might occur. Furthermore, the SIDS indices for all items in each scale (not tabulated) contain positive and negative values, which indicate cancellation across items. However, STDS and UETSDS are equal in all comparisons for the Emotional Stability and the Extraversion scale, reflecting that cancellation across items might be more present than cancellation across respondents. The differences between the STDS and UETSDS are somewhat larger for the Conscientiousness scale, which might suggest cancellation across items and across persons. However, inspecting the ETS plots for the ethnicity comparisons for the Conscientiousness scale showed that these plots are almost identical, suggesting that cancellation across persons may be very small.

Because the results indicate that cancellation is more present across items than across persons, the standardized effect size index, ETSSD, provides additional information for the practical impact of the difference. All ETSSD values are around zero, which means that the found differences are very small. Thus, the results show that despite some differences at the item level, the three different scales function similar in all three ethnic groups.

3.5 Discussion

Recently, Meade (2010) provided a taxonomy of different effect size indices for DF at the item and scale level. There is a need for effect size indices because they provide researchers an idea about the effect and practical importance of statistical significant DIF. In the present study, we applied the effect size indices proposed by Meade (2010) in two different testing situations to gain more experience with these indices.

Our results showed that although there is a motivational difference between applicants and incumbents, and as a result mean scale scores and item scores are higher for applicants than for incumbents (see also Hough, 1998, Robie, Zickar, & Schmit, 2001; Weekley, Ployhart, & Harold, 2004), this does not result in differential test functioning for the RBFP. Although LRT results showed statistically significant

DIF for all items, the effect size indices suggested that these differences were small. Our results were in agreement with the results obtained by Robie et al. (2001) and Stark et al. (2004). Robie et al. (2001) used the program DFITP4 (Raju, 1998) to investigate DIF and DTF for six scales from the Personal Preferences Inventory comparing applicants and incumbents. Their results showed that only a few items exhibit DIF and that there was no DF at the scale level.

With regard to measurement equivalence across different ethnicity groups in a selection context, we conclude that although the LRT flagged many items as statistically significant DIF, the effect size indices showed that the differences are small and also that this did not lead to DTF. Meade (2010) also showed in his cross-cultural example that DF might not be as large as studies so far indicated (e.g., Mitchelson, Wicher, LeBreton, & Craig, 2009; Sheppard, Han, Colarelli, Dai, & King, 2006).

Thus, the use of different effect size indices provides researchers a better impression about the amount of DF and their practical importance. Many items can show significant DIF, whereas differences in item and mean scores may be small. For example, when comparing applicants and incumbents, item 27 'Doubts the value of his/her personal contribution' (reverse scored) from the Emotional Stability scale exhibit highly statistical significant DIF ($G^2 = 122.6$, $df = 5$), while ESSD = .08 indicating that the effect was small.

Furthermore, although few items may function differently for different groups, the effect at the test level is often small and sometimes negligible. When comparing groups and/ or individuals across groups based on their test score, cancellation across items is appropriate. Thus, as long as the test score is of primary interest, DTF will only occur when many items in the scale exhibit uniform DIF (i.e., always favoring one group). This may, for example, occur when an item bank is used for the construction of short scales, resulting in measurement bias against one group. In this case, routinely checking effect size indices may help a researcher to get an idea about the practical importance of the differential item and test functioning.

Chapter 4
Computerized Adaptive Testing for Personality in a Business Context

4.1 Introduction

In personnel selection and career development there is a large interest in the development of computer-based tests and questionnaires (Bartram, 2000, 2006; Foxcroft, & Davies, 2006; Naglieri et al., 2004; Ployhart, Weekley, Holtz, & Kemp, 2003). Advantages of computer-based testing are increased standardization, cost effectiveness, positive image of the organization, and, in combination with the Internet, computer-based testing can ease the international recruitment and selection process (Bartram, 2006; Foxcroft, & Davies, 2006; Ployhart et al., 2003). A potential drawback is test security. Candidates can fake their identity, or may copy items and may share the content with future candidates (Bartram, 2000, 2006; Foxcroft & Davies, 2006; Schmidt, 1997). However, with the decreasing costs of personal computers and the increased networking capabilities, an increasing number of companies are using computerized testing to select their candidates.

An attractive application of the computer in personnel selection and assessment is computerized adaptive testing (CAT; Bartram, 2000, 2006; Meijer & Nering, 1999; Wiechmann, & Ryan, 2003). The foundation of CAT lies within item response theory (IRT; e.g., Embretson, & Reise, 2000; van der Linden, & Glas, 2000) modeling. In IRT the person's trait level (denoted by the Greek letter θ) and the item characteristics (such as item difficulty and item discrimination) are on a common metric. This property allows items to be individually tailored to a candidate's θ level during test administration. Another property is that once an IRT model has been fit to a pool of items a person's θ level and the standard error (SE) can be estimated using their responses to any subset of items from that pool.

The use of IRT and CAT has become popular in the ability domain (e.g., Dodd, De Ayala, & Koch, 1995; Veldkamp & van der Linden, 2002; Weiss, 2004), but also in the personality domain several applications of CAT have been discussed (Hol, Vorst, & Mellenbergh, 2001, 2005; Reise & Henson, 2000; Simms, & Clark, 2005;

Waller, 1999; Walter et al., 2007). However, the application of computerized adaptive personality testing in the business context is still rare, although CAT may have some interesting advantages. Traditionally, personality testing for assessment can be exhaustive for candidates because they are required to complete large multi-scale questionnaires. Research (Hol et. al., 2001, 2005; Reise, & Henson, 2000; Simms, & Clark, 2005; Waller, 1999; Waller & Reise, 1989) showed substantially item savings when using CAT, while maintaining a high correlation between θ estimates (denoted by $\hat{\theta}$) based on CAT and full scale $\hat{\theta}$s. Another advantage is that with the increasing use of international cross-cultural assessment, items in an item pool can be easily adapted to the requirements of different stakeholders.

The aim of the present study was to discuss the development of a computerized adaptive personality test in a business context. More specifically, the aim of the study was to apply a CAT in a real life setting and to investigate (a) the psychometric efficiency of the CAT, in particular test length reduction, time saving, and psychometric information across $\hat{\theta}$, and the correlation between CAT $\hat{\theta}$s and full scale $\hat{\theta}$s.

This study is organized as follows. First, we give a short overview of the literature on CAT and personality. Second, we describe the construction of the CAT for personality measurement in a career development context. Finally, we reflect on the advantages and disadvantages of using CAT in a business context.

4.2 CAT and Personality

In the personality domain relatively few applications of CAT exist and most of them are simulations or real data simulation studies, that is, researchers used real data to simulate CAT (Forbey, Ben-Porath & Arbisi, 2011; Gnambs& Batinic, 2010; Lei & Dai 2011; Sodano & Tracey, 2011). Waller and Reise (1989) conducted a real data CAT simulation study on the Absorption scale of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982). The item pool consisted of 34 true-false items. They applied a fixed test length and a clinical decision strategy. Using a fixed test length of 17 items, 50% item savings were achieved with little loss of measurement precision. The clinical decision strategy, which entails to administer items until the confidence interval surrounding the current trait estimate no longer includes the cutoff value used to classify persons, yielded on average 25% item savings and a perfect hit rate of individuals with extreme high Absorption scores. Waller (1999) discussed a CAT for the Denial of Somatic Complaints scale of the Minnesota Multiphasic Personality Inventory (MMPI-2; Ben-Porath & Tellegen,

2008). Waller conducted a real data CAT simulation on an item pool consisting of 51 true-false items and obtained 61% item savings.

Reise and Henson (2000) conducted a real data simulation study on the NEO Personality Inventory-Revised (NEO-PI-R; Costa, & McCrae, 1992). They administered adaptive versions for the 30 facets of the NEO-PI-R, each based on eight Likert scale items. Administration of on average four items (i.e., on average 50% item savings) resulted in a high correlation with full scale facet scores and little loss of measurement precision. The CAT algorithm resulted for most facet scales in little variability in the items selected. Therefore, Reise and Henson (2000) suggested that instead of using CAT, it might be useful to construct short forms by choosing the four "best" items that provide the highest measurement precision. However, they also acknowledged that their item pool sizes might be too small. A CAT based on a large item pool might yield different results.

In their real data simulation, Hol et al. (2001) used a Dutch version of the Dominance scale of the Adjective Check List (Gough & Heilbrun, 1980), consisting of 36 items, to study the relationship between CAT $\hat{\theta}$s and full scale $\hat{\theta}$s by manipulating the SE in the stopping rule. The authors found that $\hat{\theta}$s based on CAT were equivalent to full scale $\hat{\theta}$s. A stopping rule of SE $\leq$ .3 resulted in 22% item savings and r =.996 between CAT $\hat{\theta}$s and full scale $\hat{\theta}$s. A stopping rule of SE $\leq$ .4 yielded 67% item savings and r =.949.

Simms and Clark (2005) developed computerized adaptive versions for each of the 15 personality trait dimensions of the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993) and administered their CAT to a sample of 491 undergraduates. For each CAT, they specified a minimum number of items to be administered and the CAT terminated when either SE $\leq$ .4 or when only items that added a specified accuracy were available. Results showed 36% to 37% item savings and 58% to 60% less time to complete.

Although the studies cited above showed that CAT results in substantial item savings, most of these CAT studies consisted of simulations or real data simulation studies. In the present study, we extend the personality CAT literature by applying CAT in a real business context. According to Reise and Henson (2000) a disadvantage of real data simulations is that they are "hypothetical", that is, we do not know how persons really respond to a CAT and what this will mean for our results. Also Simms and Clark (2005) described a need for IRT based personality CATs completed by live participants. Their study was one of the first studies using live participants. However, they used undergraduates as participants.

4.3 Method

4.3.1 Development of the CAT

Item pool development

To select items for the CAT item pool, we collected data from 984 persons from the Dutch working population who were administered the Reflector Big Five Personality (RBFP; Schakel, Smid, & Jaganjac, 2007), more extensively discussed in Chapter 2, as part of a selection and assessment procedure. Participants had a mean age of 39.6 years (SD = 9.7). There were 57.7% mostly White men and 42.1% mostly White women (for .2% the gender was unknown); 24.0% had a university degree, 50.0% had higher education, and 24.4% secondary education (for 1.6% educational level was unknown).

The RBFP is a computer-based Big Five personality questionnaire applied to situations and behavior in the workplace. It consists of 144 items, distributed over five scales (Need for Stability, Extraversion, Openness, Agreeableness, and Conscientiousness). The items are scored on a five point Likert scale. The answer most indicative for the trait being measured is scored '5' and the answer least indicative for the trait is scored '1'. The questionnaire is administered online. Coefficient alpha varies from .79 to .91 for the five scales.

Items from the three Big Five factors Need for Stability[2], Extraversion, and Conscientiousness were selected for the CAT item pool. For this study, we recoded the Need for stability scale such that it can be interpreted as an Emotional Stability scale. In that way, all three scales are phrased in a positive direction. Based on the Company's own selection and assessment experience, they were most interested in measuring these three factors. From these original three scales, items were selected that allowed unidimensional measurement and discriminated well between persons.

To select items that together formed a scale, we checked the assumptions of the Mokken model of monotone homogeneity (MMH, e.g., Sijtsma & Molenaar, 2002) using the computer program Mokken Scale Analysis for Polytomous Items version 5.0 for Windows (MSP5.0, Molenaar & Sijtsma, 2000) by inspecting the H and H_g coefficients and by inspecting the item step response functions (ISRFs). In the Appendix, we discuss the ISRF in more detail. Under the MMH the ISRFs are monotone nondecreasing functions in θ. We use the coefficient H_i for items

[2] This itempool is a little bit larger than the original scale of the RBFP.

($i = 1, \ldots k$) and coefficient H for a set of items. Increasing values of H and H_i between .30 and 1.00 (maximum) mean that the evidence for monotone increasing ISRFs is more convincing, whereas values below .30 indicate violations of increasing ISRFs (for a discussion of these measures see for example, Meijer & Baneke, 2004 or Sijtsma & Molenaar, 2002). Furthermore, weak scalability is obtained if $.30 \leq H < .40$, medium scalability if $.40 \leq H < .50$, and strong scalability if $.50 \leq H < 1$ (Sijtsma & Molenaar, 2002, pp. 60-61). Because an initial analysis showed that a number of items had H_i values just below $H_i = .30$, and because we did not want to throw away possible useful items, we used a liberal lower bound of $H_i \geq .25$ to select items in the item pool. This resulted in an item pool of 31 Emotional Stability items ($H = .31$), 27 Extraversion items ($H = .34$), and 23 Conscientiousness items ($H = .33$). Further inspection of the ISRFs showed no significant violations against monotone nondecreasing ISRFs.

CAT selection algorithm

Item parameters were estimated using the graded response model (GRM; Samejima, 1969, 1997) and the computer program MULTILOG (Thissen, 2003) with marginal maximum likelihood (MML) estimation. The GRM is a 2-parameter-logistic (2PL) model for polytomous data. The model assumes that a person makes a global evaluation before responding to an item. For example, for an item with four categories, the person compares the first category with the second, third and fourth category; the first and second with the third and fourth category; and the first, second and third category with the fourth category. The person immediately seeks his/her position on the scale. Each item i is described by one item slope parameter (a_i; 'item discrimination') and $j = 1, \ldots, m_i$ between category "threshold" parameters (b_{ij}). We denote $m_i + 1 = K_i$ to be equal to the number of item response categories within an item. These parameters are used to determine the probability of an examinee to respond in a particular response category (x) (Embretson, & Reise, 2000). Consider an item with K = 5 response categories, then there are $m_i = 5 - 1 = 4$ thresholds. The probability that a candidate responds in category x ($x = 0 \ldots 4$) conditional on θ equals:

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i(x+1)}(\theta)$$

with

$$P^*_{ix}(\theta) = \frac{e^{a_i(\theta - b_{ij})}}{1 + e^{a_i(\theta - b_{ij})}}$$

and $x = j = 0 \ldots 4$, $P^*_{i0} = 1.0$ and $P^*_{i4} = 0.0$

Important tools in the context of CAT are the item and test information curves (Embretson, & Reise, 2000). The item information curve indicates the amount of psychometric information an item provides at each level of $\hat{\theta}$, based on the *a* parameter and the probabilities of responding in a certain category. These curves are additive across items on a common scale and together constitute the test information ($TI(\hat{\theta})$). Test information indicates the amount of information a test provides at each level of $\hat{\theta}$. It has an exact relationship with a candidate's standard error of measurement, namely,

$$SE(\hat{\theta}) = \frac{1}{\sqrt{TI(\hat{\theta})}}.$$

We used the "best guess method" for the initial item selection, that is, the CAT started with an item of medium difficulty (Parshall, Spray, Kalohn, & Davey, 2002). To select a next item in the CAT, the item with the highest amount of information at the candidate's current trait level was selected (Parshal et al., 2002). As a stopping rule we used $SE < .32$. This corresponds to a reliability of $\varrho \geq .9$ for each individual (Daniel, 1999, p. 54). However, preliminary analyses revealed that for persons at the right side of the θ continuum it was difficult to reach $SE < .32$. The reason was that our item pool contained too few items that discriminated well between highly emotional stable, extravert, or conscientious persons. This is illustrated in Figure 4.1 (to be discussed in more detail below) where the $TI(\hat{\theta})$s for the three personality scales are depicted. Consequently, for these persons we could not obtain enough information to reliably estimate their θs. Therefore, we also stopped the CAT when the increase in item information was less than .25. Finally, to avoid that the CAT is finished after, for example, only four items, we set the minimum number of items administered at nine items for each person.

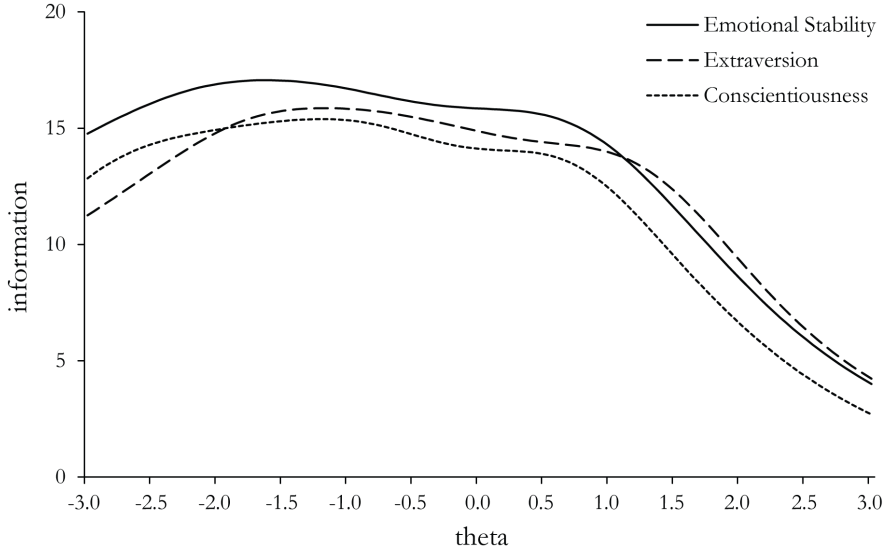


Figure 4.1: Test information curves for the CAT scales.

4.3.2 Participants and Procedure

Data were collected as part of a career development procedure in the context of a reorganization of a Dutch professional care company. Based on their personality scores and interviews with a human resource manager, participants were selected for a particular vacancy. There were a few internal relocations, but mostly it concerned outplacement. Because the organization did not want to put much pressure on their employees, the employees did not have to fill out an intelligence test. There were 428 participants with a mean age of 32.1 (*SD* = 12.7); 29% mostly White men and 71% mostly White women. 28.7% of the participants had a university degree, 59.1% had higher education, and 12.1% secondary education.

Data were collected in an unproctored setting; participants received an email with instructions and a hyperlink to the CAT. Eight days later they received an email with a hyperlink to the RBFP. Consequently, it was possible to compare the score on the CAT scales with a full scale score based on the RBFP answers. After finishing the RBFP, participants received a report in their mailbox with their individual scores on the different factors and facets, together with the interpretation of their scores. This report was used as input for further interviews with their human resource manager.

Table 4.1

Item savings, reliability and mean psychometric information at the scale and item level

	mean number of items			mean psychometric information per scale			alpha			mean psychometric information per item		
Scale	full	CAT	%sav	full	CAT	%loss	full	CAT	%loss	full	CAT	RE
Emotional Stability	31	15.2	51	15.7	9.7	38	.94	.92	2.1	.51	.66	1.30
Extraversion	27	13.5	50	14.9	10.2	32	.94	.92	2.1	.55	.77	1.39
Conscientiousness	23	12.9	44	14.1	10.0	29	.91	.89	2.2	.61	.78	1.28
Overall	81	41.7	49	44.7	29.8	33	NA	NA	NA	.56	.74	1.32

Note. alpha = 1- MSE^2/SD^2 (Daniel, 1999, p. 54), NA = not applicable, $RE_{CAT/full}$ = relative efficiency between CAT and full scale administration

4.4 Results

4.4.1 CAT and Full Scale Comparison

Measurement Precision and Efficiency

Table 4.1 displays the item savings, reliability, and mean psychometric information at the scale and item level, when comparing CAT and full scale administration. The overall item saving equalled 49%. The Emotional Stability scale has the largest item savings with 51%. As a consequence of these item savings, there are differences in mean psychometric information at the scale level. Overall there was a loss of psychometric information of 33%, when comparing the CAT scores with the full scale scores suggesting that the full scale scores are more precise than the CAT scores. However, this loss of psychometric information is equivalent to a decrease in reliability of only .02 points for all three scales.

When we consider the mean psychometric information per item (i.e., the psychometric efficiency), the CAT has a higher efficiency per administered item than the full scale. The relative efficiency (i.e., the CAT efficiency divided by full scale efficiency; Hambleton, Swaminathan, & Rogers, 1991) for the CAT equalled 1.32. For the Extraversion scale the relative efficiency was largest (1.39). This implies that although CAT administration resulted in loss of psychometric information in an absolute sense, the CAT is approximately 30-40% more efficient than full scale administration in terms of information per item administered.

Administration Time

Besides efficiency based on item savings and psychometric information, we investigated the reduction in administration time for the CAT as compared to the full scale. CAT yielded a reduction in administration time of 52.8%, which means, on average, a reduction of more than five minutes. This reduction is a significant decrease; a paired *t* test for the full scale (M = 596s, SD = 202) and CAT (M = 281s, SD = 78) comparison yielded $t(427) = 41.65$ ($p < .001$). Simms and Clark (2005) found a reduction in administration time of 38%, which corresponded to a reduction of almost 8 minutes related to the almost 20 minutes to complete the full scale of 304 items in their study.

The administration time per item for the full scale (M = 7.36s, SD = 2.49) and CAT (M = 6.78s, SD = 1.72) yielded a significant difference in time per item; $t(427) = 7.42$ ($p < .001$). Thus, participants needed less administration time to answer an

Table 4.2

Response frequencies and administration order for the items in the Conscientiousness scale.

	Item number (*a* parameter)																							
Admin order	1 (1.25)	2 (1.69)	3 (1.11)	4 (1.51)	5 (1.11)	6 (1.52)	7 (1.31)	8 (1.63)	9 (1.88)	10 (2.00)	11 (2.10)	12 (1.20)	13 (1.28)	14 (1.13)	15 (1.42)	16 (1.45)	17 (1.18)	18 (1.17)	19 (1.91)	20 (1.02)	21 (1.99)	22 (1.50)	23 (1.55)
1st									85	84	85								90		84		
2nd										166	114			129							11	8	
3rd			23	58			1			63	159								31		93		
4th				19		1			122	47	40				15				80	8	96		
5th		8		18					128	37	6	4			7	23			119		78		
6th		272	4	2	4			16	35	7	4	3			10	24			22		23		2
7th		67		2		3		261	9	3	5	14			13	8			14		13		16
8th		30	4	147		10		32	10	2	2	9	2		4	3	5		10	9	5		144
9th		11	3	126		25		25	4	1	1	4	2		5	15	13		5	2	3	5	178
10th		7		37		110		9	5		1	1	3		12	133	4		5	1	2	29	53
11th		3	2	18		128	1	6	1	1			3		40	142	1		4	1	2	46	10
12th		5	1			46		9	1	1		5	8		155	48			10		1	32	4
13th	1	1				19		16	1		1	5	13		59	12	1		5	2		90	1
14th					1	18		10	3			5	23	2	9		8		2		1	3	
15th		1	2		1	11		5	1			11	15	1			6	1	6			1	
16th		1	3		3	8		7	1			13	1	7			4			2		1	
17th			5		3	9		1				1			1		18					1	
18th			17		3	3		3				1					3		2	2			
19th					19	1		1								1			2	2			
20th				1																18			

item in the CAT administration mode, compared to the full scale mode. This may be due to the CAT selection algorithm that selects the items that are most indicative for the candidate's θ at that time in the CAT.

Relation between CAT and Full Scale

Correlations between CAT and full scale scores were determined; Emotional Stability, Extraversion, and Conscientiousness CAT scores correlated .83, .88 and .84, with the full scale scores, respectively. Correlations between CAT scores and scores on the original RBFP counterparts were also computed; CAT scores correlated .73, .74 and .76 with their RBFP counterparts for the Emotional Stability, Extraversion, and Conscientiousness scale, respectively. These results indicate that besides time savings and substantially item savings when using CAT, high correlations between CAT and full scale scores and CAT and the original RBFP scale scores were obtained.

4.4.2 Item Presentation Analyses

We evaluated the item administration order and whether candidates with different $\hat{\theta}$s were receiving different CATs. Reise and Henson (2000) found that the item administration order demonstrated little variability across candidates and that choosing the 'best' items, that is, the items with the highest item information, resulted in similar results as CAT administration.

We present a brief overview of the most important results. First, we correlated for each scale the *a* parameters with the number of times the item was administered and with the serial position in the CAT administration. The *a* parameters correlated significantly with the number of times an item was administered (r = .84, .85 and .83 for the Emotional Stability, Extraversion, and Conscientiousness scale, respectively). The *a* parameters also correlated highly with the serial position of an item (r = -.53, r = -.57 and r = -.76, for the Emotional Stability, Extraversion, and Conscientiousness scale, respectively); highly discriminating items were more likely to be administered at the beginning of the CAT.

However, these results do not imply that all candidates received the same CAT, although there is some overlap. To illustrate this, in Table 4.2 the item administration order is given for the Conscientiousness scale. At each stage of the CAT one or two items are administered to a large amount of the candidates, but almost never to more than 50% of the candidates. Item 2 at the 6th stage and item 8 at the 7th stage are exceptions; these items were administered to 64% and 61% of

Table 4.3

Item parameters of the Emotional Stability scale and the percentage an item is selected and the serial position in the CAT for each $\hat{\theta}$ category

	item parameters					percentage selected in CAT			serial position in CAT		
	a	b_1	b_2	b_3	b_4	cat 1	cat 2	cat 3	cat 1	cat 2	cat 3
item 1	1.15	-2.84	-.73	.46	2.02		9	100		15.0	9.7
item 2	1.08	-2.01	.21	1.40	3.17		15	98		3.7	6.1
item 3	1.75	-2.04	-.87	-.30	1.09	98	100	83	4.5	3.1	4.6
item 4	2.05	-3.49	-2.28	-1.31	.60	100	100	78	2.2	3.0	6.2
item 5	.91	-2.67	-.07	1.09	3.40		2	43		4.0	5.0
item 6	1.10	-4.32	-2.37	-1.19	.62	2			17.0		
item 7	1.25	-2.66	-1.04	-.20	1.73		35	93		14.2	9.3
item 8	1.42	-3.22	-1.42	-.56	1.07	97	100	73	11.2	11.0	10.3
item 9	1.88	-2.02	-.98	-.35	.81	100	100	77	3.2	2.9	5.9
item 10	1.53	-3.24	-1.84	-.95	.47	100	100	50	7.9	8.8	17.4
item 11	1.23	-4.06	-2.16	-.87	1.38	2	12	75	8.0	15.9	13.6
item 12	1.21	-3.79	-2.78	-1.84	-.14	9			14.4		
item 13	.83	-2.21	.12	1.51	4.01	5	31	87	2.0	2.0	2.0
item 14	1.44	-2.58	-.72	.08	1.60	64	100	90	12.7	9.2	5.4
item 15	1.46	-5.13	-2.84	-1.91	.46	53	71	38	10.0	13.7	20.7
item 16	1.50	-4.10	-2.00	-.71	1.19	100	100	77	9.2	9.7	8.6
item 17	1.67	-3.22	-1.58	-.79	1.14	100	100	78	6.6	6.9	6.6
item 18	1.69	-2.87	-1.73	-1.09	.50	100	100	58	5.6	6.6	12.6
item 19	.99	-3.80	-2.38	-.66	1.66			23			21.7
item 20	1.21	-3.87	-2.06	-.55	1.53	2	9	82	12.0	16.7	12.7
item 21	1.38	-2.66	-1.50	-.76	.72	100	100	65	11.7	12.5	15.1
item 22	1.23	-4.76	-3.27	-1.70	.60	3		7	11.0		24.0
item 23	1.72	-3.05	-1.80	-1.12	.43	100	100	63	3.3	4.7	9.3
item 24	1.16	-3.14	-1.83	-.35	1.38	2	4	72	18.0	18.7	16.5
item 25	1.03	-2.56	-1.04	.35	1.97			85		8.0	15.6
item 26	1.04	-3.31	-1.28	-.20	1.87			72			17.2
item 27	1.05	-2.76	-1.28	-.45	1.38			42			22.5
item 28	1.39	-3.27	-2.11	-1.30	.59	98	96	58	11.3	13.9	19.1
item 29	1.03	-4.19	-1.64	-.30	2.04		1	77		5.5	14.5
item 30	1.24	-4.19	-2.44	-1.48	.81	3	2	62	11.5	17.8	20.5
item 31	1.78	-3.77	-2.54	-1.82	-.07	100	94	22	4.2	5.2	1.0

Note: cat 1 = $\hat{\theta}$ smaller than 1SD below the mean; cat 2 = $\hat{\theta}$ between 1SD below and above the mean; cat 3 = $\hat{\theta}$ larger than 1SD above the mean.

the participants, respectively. This indicates that although there are some items that dominate a CAT stage, different items are selected at each stage. Note that items 1, 7, and 18 were administered only a few times due to their relatively low *a* parameters. Item 4 is the only item that is selected for all candidates. This may be due to a broad range of the *b* parameters and a relatively high *a* parameter. The five items that were selected most often (items 4, 11, 10, 16, and 19) are not the five items with the highest *a* parameters. Similar results were found for the other two scales.

We also investigated if the number of items selected in the CAT differed across $\hat{\theta}$ and if candidates with different $\hat{\theta}$s were receiving different CATs, thus different items. Therefore, we grouped the persons according to their $\hat{\theta}$ values for each scale. The first category ranged from the lowest $\hat{\theta}$ through the $\hat{\theta}$ value that is one standard deviation below the mean $\hat{\theta}$. The second category contained the $\hat{\theta}$s between minus one and plus one standard deviation around the mean and the third category contained the $\hat{\theta}$ values larger than one standard deviation above the mean. Less emotional stable candidates (first $\hat{\theta}$ category) received, on average, 13.4 items, moderately emotional stable candidates (second $\hat{\theta}$ category) received, on average, 14.8 items, and highly emotional stable persons (third $\hat{\theta}$ category) received on average, 19.3 items. So, the number of items selected in the CAT differed across $\hat{\theta}$.

For the Emotional Stability scale, Table 4.3 displays the item parameters, the percentage an item is selected and the serial position of an item in the CAT for each $\hat{\theta}$ category. Some items, like item 3 "is afraid of making mistakes" and item 17 "recovers promptly after setbacks" are often selected. These items have a broad range of *b* parameters and high *a* parameters. However, these items differ in the serial position in the CAT. Item 3, for example, is selected earlier in the CAT for the moderate emotional stable candidates than for the highly emotional stable candidates.

Interestingly, there are also items that are (almost) only selected for a particular $\hat{\theta}$ category, especially for the third $\hat{\theta}$ category. For example item 1 "takes criticism personally" is selected for all candidates with $\hat{\theta}$s larger than one SD above the mean $\hat{\theta}$. Because most items are situated at the left side and in the middle of the trait continuum, a number of items are always selected for candidates with $\hat{\theta}$ ranging from the lowest through $\hat{\theta}$ one SD above the mean although the serial position differs for most items. For example item 10 "sees problems rather than solutions" is always selected for candidates in the first and second $\hat{\theta}$ category and only for 50% of the candidates in the third category (i.e., the highly emotional stable candidates).

Table 4.4
Item parameters of the Extraversion scale.

	a	b_1	b_2	b_3	b_4
item 1	1.57	-3.21	-2.13	-1.24	0.51
item 2	1.87	-1.68	-0.60	0.00	1.09
item 3	0.97	-2.27	-0.56	0.78	3.22
item 4	1.86	-2.46	-0.99	-0.27	1.48
item 5	1.31	-3.63	-1.89	-0.87	1.06
item 6	0.94	-5.37	-3.42	-1.04	1.66
item 7	1.51	-3.43	-1.86	-0.98	0.76
item 8	1.35	-2.99	-1.65	-0.52	0.92
item 9	1.72	-2.04	-1.20	-0.56	0.53
item 10	2.15	-1.60	-0.71	-0.09	1.00
item 11	1.40	-1.67	-0.29	1.23	2.89
item 12	1.64	-3.55	-2.38	-1.26	0.85
item 13	1.95	-1.76	-0.88	-0.02	1.54
item 14	1.32	-2.43	-0.82	-0.15	1.84
item 15	0.95	-2.54	-0.32	0.78	3.18
item 16	0.96	-3.10	-0.74	0.43	2.96
item 17	1.25	-4.25	-2.01	-1.04	1.15
item 18	1.75	-3.30	-1.71	-0.80	1.35
item 19	1.19	-6.21	-3.95	-2.31	-0.07
item 20	1.30	-4.73	-3.05	-1.11	1.46
item 21	1.28	-5.30	-2.49	-1.00	1.30
item 22	1.24	-7.24	-4.15	-2.34	0.50
item 23	1.32	-3.17	-1.71	0.01	1.81
item 24	0.90	-2.28	-0.30	0.68	3.05
item 25	1.47	-2.92	-1.48	-0.69	0.99
item 26	1.14	-5.84	-2.60	-1.45	1.12
item 27	0.87	-2.28	-1.03	0.05	2.25

4.4.3 Accuracy across $\hat{\theta}$

Inspecting the test information functions for each scale in Figure 4.1, we obtain information about the measurement accuracy across $\hat{\theta}$. From Figure 4.1 it is clear that the item pool is especially suited to reliably distinguish candidates who score in the middle or the left end of the trait continuum. From a personnel selection and assessment perspective this makes sense. These scales are especially developed to reliably distinguish neurotic, introvert, and disordered candidates from emotional stable, extravert, or conscientious candidates, respectively. However, the item pool is less suited to distinguish, for example, moderate extraverts from extreme extraverts. When selecting candidates for jobs that require extreme extraverts this may be important. Examples are salesmen who have to convince strangers to buy their products (Furnham & Fudge, 2008) or special agents and spies who have to be extremely good at socializing and striking up conversations with strangers because part of their job is to recruit foreign assets (Waller, 1993). Personality questionnaires used to select these types of candidates thus should contain many items that tap into the high end of the Extraversion scale.

To illustrate this, assume that we want to select persons that are one standard deviation above the mean score on the Extraversion scale, then there are only a few items in our item pool that are suitable to measure these persons accurately. Inspecting Table 4.4 shows that even the third threshold parameter (response categories 1, 2, 3, versus 4, 5) was in the negative range for 19 out of the 27 items. Thus, for example for item 7 ("makes the first move for face-to-face contact") the third threshold parameters was b_3 = -.98, implying that even persons who are one standard deviation below the mean on the Extraversion scale are most likely to respond in the two highest categories.

The question is whether we can construct items that tap into the right end of the Extraversion continuum. In our item pool, Item 11 "likes to have others lead meetings" (reverse-scored) is the item that gives the most information at the right side of the Extraversion scale. Candidates with a θ value one standard deviation above the mean are most likely to respond in the highest categories (b_3 = 1.23 and b_4 = 2.89). In the literature (e.g., Reise & Henson, 2003) there is some discussion whether it is possible to construct items that can be used to measure extreme latent trait values. A lot of researchers are operating under the assumption that all constructs are fully continuous, defined at both ends of the construct, and that items can be found that measure accurately across the entire range. It is an empirical question whether this is possible. Another problem is that in a personnel selection

context social desirable answering may play a role. For example, when applying for a sales job not many candidates will answer negatively to the question "makes the first move to face-to face contact".

4.5 Discussion

The aim of the present study was to extend the personality CAT literature by applying a CAT in real selection and assessment practice using real participants. The results indicated that CAT administration resulted in a reduction of approximately 50% of the number of items and the testing time as compared to a full scale questionnaire. These item savings were expected given previous IRT based CAT studies (Hol, et. al., 2001, 2005; Reise, & Henson, 2000; Simms, & Clark, 2005; Waller, 1999; Waller, & Reise, 1989; Weiss, 2004). The CAT version resulted in a loss of psychometric information of approximately 33%, suggesting that the full scale scores are more precise than the CAT scores. In contrast to the Reise and Henson (2000) study, we found that although item discrimination clearly predicted when and how many items were selected to the candidates, we also found that candidates received different CATs. This may be explained by the larger item pool we used compared to the item pool in the Reise and Henson (2000) study.

Industrial and Organizational psychologists and recruiters typically administer personality questionnaires as a way to communicate about a person. In practice, self-report personality questionnaires are almost never the only source of information available to a psychologist on the basis of which decisions are based. Self-reports are almost always combined with cognitive measures and data from interviews. Nevertheless, it may save time and energy to administer a personality questionnaire as efficient and effective as possible. Through the construction of a CAT we also learned that our item pool is less suitable for highly emotional stable, extravert, and conscientious persons. For these persons, there were not enough items to reliably estimate their θs.

4.6 Appendix

An item step is the imaginary threshold between adjacent ordered response categories. As an example, imagine the personality item "Seldom experiences a feeling of failure" having three ordered answer categories (disagree, agree, strongly agree). It is assumed that the participant first ascertains whether he or she agrees enough with the statement to take the first item step (from disagree to agree). If not, the first item step equals 0, and the item score also equals 0. If the answer is affirmative, the item step equals 1, and the participant has to ascertain whether the second step (from agree to strongly agree) can be taken. If not, the second item step equals 0, and the item score equals 1. If the answer is affirmative, the second item step score equals 1, and the item score equals 2. The ISRF describes the relation between the probability that the item step score equals 1 and θ. An item with three ordered answer categories has two item steps and consequently, two ISRFs, one for each item step. The MMH assumes that each of the ISRFs is monotone nondecreasing in θ.

Chapter 5
Invariant Item Ordering and the Reflector Big Five Personality

5.1 Introduction

As discussed in Chapter 1 many organizations now publish their job vacancies on the Internet, as this is a cheap and efficient way of bringing them to the attention of prospective candidates. This has three major ramifications for the use of tests and questionnaires: the first is that questionnaires are completed in an unproctored setting, the second is that candidates' test and questionnaire results can be used to determine their suitability for several jobs with different job descriptions and the third is the demand for short tests and questionnaires. In Chapter 6 we will deal with methods to check whether results from an unproctored online test can be compared with the results on a proctored one. In this chapter we discuss the possibility to select subsets of items from an item bank that have the same "difficulty" order for low and high scoring persons, which might result in short tests.

When items are stored in an item bank or portal, it may ease the construction of new instruments or the use of subsets of items from an existing instrument when an assessor knows that a particular personality item reflects an extreme point-of-view for every candidate, irrespective of a candidate's trait value. For example, in a selection context when measuring emotional stability, it may be very informative to know that the items have a similar ordering for persons with different latent trait values. In this context the item "faith in own ability to tackle problems" is expected to be endorsed more often than an item "needs no confirmation from others". It is often assumed that the item ordering according to severity (or mean score) established at the *group* level is the same for persons at different *individual* trait levels. However, as Ligtvoet, van der Ark, te Marvelde, and Sijtsma (2010) and Sijtsma, Meijer, and van der Ark (2011) discussed, this assumption only holds when the items form a hierarchical scale. Items form a hierarchical scale when the ordering of the items according to their severity is the same across different values of the latent variable. This property is named invariant item ordering (IIO).

IIO is especially useful when a researcher or a psychologist is comparing or diagnosing individual persons. For example, IIO facilitates the comparison of children with respect to the development in transitive reasoning (e.g., Bouwmeester & Sijtsma, 2006), but it may also facilitate diagnosing personality trait scores. For example, Watson, Deary, and Austin (2007) investigated whether the items of the Neurotocism scale of the NEO-FFI formed a hierarchical scale. Recently, Meijer and Egberink (in press) analyzed different clinical scales and found that for these scales many items did not comply to the property of IIO. One of the reasons was that many items were replications of each other.

As discussed in Sijtsma et al. (2011) in clinical, health, and personality measurement checking for IIO may have the following important advantage: "Let us conceive of items as symptoms; then, when IIO holds, compared to a person with a lower score, a person with a higher score has the same symptoms plus more symptoms representing higher intensity levels. This hierarchy of symptoms can be inferred from the total score and supports the useful interpretation of total scores, not only as indicators of attribute levels but also as summaries of particular sets of symptoms. The higher the total score, the more the set of symptoms is extended with additional ones, and symptoms are always added in the same order" (p. 32).

Sets of items in clinical, health, and personality inventories are seldom checked on the IIO property and if they are, often suboptimal methods are being used (Meijer, 2010). As discussed in Sijtsma et al. (2011) only a few models allow for IIO. For polytomously scored items, only a few restrictive polytomous IRT models imply IIO (Sijtsma & Hemker, 1998), such as the rating scale model (Andrich, 1978) and a rating scale version of Muraki's (1990) restricted graded response model.

The aim of the present chapter is to check whether the items of the short version of the Reflector Big Five Personality (RPBF; Schakel, Smid, & Jaganja, 2007) formed a hierarchical scale. Doing this, we obtain insight into (1) the usefulness of methods that have been proposed to establish IIO, (2) the psychometric quality of the RPBF, and (3) the sometimes difficult decision when to remove an item from a scale because it violates the IIO property. In particular this last issue has received remarkable little attention in the literature, whereas it plays a crucial role in scale construction. It is important to stress that we do *not* want to advocate that every test or scale should have the IIO property. Instead, we would like to show that investigating IIO may help a researcher to obtain a better understanding of psychological item scores and test scores.

Because IIO has been formulated in the context of item response theory (IRT; Embretson & Reise, 2000), we first discuss two nonparametric IRT models that are

relevant in this context. Second, we discuss different methods to establish IIO. Third, we illustrate how IIO can be empirically investigated, and finally we discuss the practical consequences for personality measurement.

5.2 Nonparametric Item Response Theory

In the present study, we follow a nonparametric IRT approach to investigate IIO for which two nonparametric IRT models are relevant: Mokken's model of monotone homogeneity (MMH; Mokken, 1971, 1997) and Mokken's double monotonicity model (DMM; Mokken, 1971, 1997). The MMH assumes increasing item response functions (IRFs). The IRF denotes the probability that an item i is answered correctly or is endorsed in the keyed direction for a specified value of the latent trait θ and is denoted $P_i(\theta)$. Nonparametric and parametric IRT models differ with respect to the form of the IRF. In nonparametric models there are no restrictions with regard to the form of the IRFs, except that they should be increasing. The DMM also assumes increasing IRFs, but an additional assumption is that the IRFs do not intersect. This makes the DMM a special case of the MMH, which means that when the DMM holds the weaker MMH also holds, but the reverse is not true. The assumption of nonintersecting IRFs implies IIO. More formally, when IIO holds for a set of k items and the items are ordered to decreasing popularity (or decreasing proportion correct score), it applies that

$$P_1(\theta) \geq P_2(\theta) \geq \ldots, \geq P_k(\theta), \text{ for all } \theta. \quad (1)$$

Molenaar (1997) discussed polytomous versions of Mokken's original dichotomous models, which are based on the same set of assumptions as the MMH model. Central in his approach is the item step response function (ISRF). Let X_i be the score on item i, with values $x_i = 0, \ldots, m$; for 5-point rating scales, this means $x_i = 0, \ldots, 4$. The ISRF is the probability of obtaining an item score of at least x_i and is denoted $P(X_i \geq x_i \mid \theta)$ for $x_i = 1, \ldots, m$, thus ignoring $x_i = 0$ because this probability by definition equals 1. Molenaar (1997) also discussed the DMM for polytomous items, which adds to the MMH the assumption that the ISRFs of different items do not intersect.

The polytomous DMM model, however, does *not* imply that items can be invariantly ordered. This has been extensively discussed in Sijtsma et al. (2011) and Meijer (2010), but has been a source of confusion in a number of empirical papers. For example, Watson et al. (2007), Watson, Roberts, Gow, and Deary (2008), Diesfeldt (2004), and Rivas, Bersabé, and Berrocal (2005) claimed to investigate whether sets of items have IIO. In all these studies, however, methods were used

that are sensitive to checking whether sets of ISRFs do not intersect, not whether items have IIO. See Roorda et al. (2005) for a good example how to investigate IIO.

5.3 Methods to investigate IIO

Several methods have been developed to establish IIO for dichotomously and polytomously scored items (Sijtsma & Junker, 1996, for an overview). We restrict ourselves to methods for polytomously scored items.

5.3.1 Method Manifest IIO

Ligtvoet et al. (2010) developed a method to investigate IIO for polytomous items, which is named method manifest IIO. Method manifest IIO compares the ordering of the item means for all item pairs for different rest-score groups, with again, the rest score, $R_{(ij)}$, as the total score on $k - 2$, thus without the scores on items i and j. IIO holds when

$E(X_j \mid R_{(ij)} = r) \geq E(X_i \mid R_{(ij)} = r)$, for all r and all item pairs.

This is investigated by numbering and ordering the items according to their conditional sample mean scores for all r. Then, a one-sided one-sample t-test is conducted to test the null hypothesis that the expected conditional item means are equal against the alternative that the expected conditional mean of an item i exceeds that of item j, which is a violation of IIO. A violation is reported when there is a reverse ordering of the conditional sample means for a particular rest score. To prevent that very small violations are taken seriously, these reverse orderings are only tested when they exceed a minimum value, denoted *minvi*.

5.3.2 Coefficient H^T

Coefficient H^T (Ligtvoet, et al., 2010) can be used as a measure for the accuracy of the item ordering. A low H^T value suggests that the IRFs are close together, whereas a high value of H^T suggests that the IRFs are further apart. When IIO holds for k items, it can be shown that $0 \leq H^T \leq 1$. For practical purposes, Sijtsma and Meijer (1992) suggested to use $H^T \geq .3$ as a lower bound. It is important to emphasize that H^T is only related to all k items together, and cannot be used to assess which items cause intersections.

5.4 Method

5.4.1 Instruments and Data

In this study, we used a short version of the RBFP (hereafter referred to as RBFP; Schakel, et al., 2007). As discussed in Chapter 2, the RBFP is an online computer-based Big Five personality questionnaire applied to situations and behavior in the workplace. The short version of the questionnaire is used as a global assessment of the Big Five factors. It consists of 72 items, distributed over five scales (Need for Stability, Extraversion, Openness, Agreeableness, and Conscientiousness). The items are scored on a five point Likert scale. The answer most indicative for the trait being measured is scored '4' and the answer least indicative for the trait is scored '0'. This short version of the RBFP is also based on the Workplace Big Five Profile constructed by Howard and Howard (2001), which is based on the NEO-PI-R and adapted to workplace situations. Data were collected by the Company whenever a personality measure was administered to a client. The participants were employed at an organization and completed the short version of the RBFP as part of their own personal career development. The sample consisted of 1444 persons (M_{age} = 39.7, *SD* = 9.27); 54.4% men and most persons were White. 26.0% of the participants had a university degree, 52.1% had higher education, and 21.8% secondary education. Based on a first analysis (classical indices are given in Table 5.1), we selected the subscales Emotional Stability[3] and Conscientiousness for the IIO analyses. These two scales had a relatively high estimated reliability and relatively large variation in item means.

Table 5.1
Cronbach's alpha, the range of the item-test correlations, and the range of the item means for the fives subscales of the RBFP.

Scale	nummer of items	alpha	range item-test correlations	range item means
Emotional Stability	15	.85	.32-.61 (.08)	1.74-3.56 (.50)
Extraversion	15	.80	.20-.59 (.12)	2.23-3.48 (.32)
Openness	12	.84	.32-.68 (.12)	2.42-3.46 (.31)
Agreeableness	15	.58	.02-.33 (.08)	.90-3.36 (.83)
Conscientiousness	15	.81	.21-.59 (.12)	2.02-3.58 (.48)

[3] For this study, we recoded the Need for Stability scale such that it can be interpreted as an Emotional Stability scale.

5.4.2 Data-analyses: Investigating IIO

Before investigating IIO, we ran the option TEST in MSP5 (Molenaar & Sijtsma, 2000). This option can be used to obtain insight into the psychometric quality of the total scale. We were especially interested in coefficient H_i for items and the coefficient H for a set of items. Under the MMH, higher positive H_i values reflect higher discrimination power of the items, and as a result, more confidence in the ordering of respondents by means of their total scores. For practical test construction purposes, the following rules of thumb have been suggested. Weak scalability is obtained if $.3 \leq H < .4$ medium scalability if $.4 \leq H < .5$ and strong scalability if $H \geq .5$ (Sijtsma & Molenaar, 2002). Values of H smaller than .3 are considered evidence that the items are unscalable for practical purposes.

To investigate IIO we followed the methodology described by Sijtsma et al. (2011). For polytomous items they distinguished the following steps (1) investigate overall scale quality through an automated item selection procedure (AISP), (2) investigate monotonicity through inspecting item rest-score regressions, (3) investigate IIO through the method manifest IIO proposed by Ligtvoet et al. (2010), and (4) investigate the precision of the item ordering through the H^T coefficient. These analyses were performed using the *R* package mokken (Van der Ark, 2007).

The AISP aims at selecting items from a given item pool that satisfy particular scaling criteria. The procedure starts with selecting the item pair with the highest item pair H_{ij} value significantly larger than 0 and exceeds a lower bound c. Then a third item is chosen that (a) correlates positively with the first two items, (b) has an H_i coefficient that is larger than lower bound c, and (c) results in the largest H_{ij} value for the selected items. This procedure continues selecting a next item from the item pool until criteria cannot be met.

Because significant violations of monotonicity sometimes have low impact, Molenaar and Sijtsma (2000) discussed an effect size measure named *Crit* that consists of a weighted number of different indicators of violations for which the following guidelines have been suggested: *Crit* values smaller than 40 indicate no serious violations; *Crit* values between 40 and 80 indicate minor violations, and *Crit* values larger than 80 indicate serious violations. We used these *Crit* values to get an idea about the seriousness of model violations.

Ligtvoet et al. (2010) suggested the following sequential data-analysis procedure for method manifest IIO. First, for each of the k items the number of significant violations (i.e., that exceed *minvi*) is determined and the item with the highest number of violations is removed. When different items have the same number of

significant violations, the item with the smallest H_i coefficient may be removed, but also other criteria might be considered, for example the item content. Second, this procedure is repeated for the remaining items until none of the items have significant violations, which means that IIO holds for all k items. When IIO holds for the (remaining) k items, the H^T coefficient for polytomous items can be computed, which is a generalization of the H^T coefficient for dichotomous data to obtain an idea about the accuracy of the item ordering.

The analyses were conducted using default settings in both programs, that is, we used a lower bound of c = .30 for the AISP, *minvi* = .03 to investigate monotonicity, and *minvi* = .03 times the number of item step response functions (i.e., *m)* to investigate IIO. Ligtvoet et al. (2010) investigated the sensitivity and specificity of method manifest IIO. They used different *minvi* values and their simulation study showed that a *minvi* of .03 times *m* is an appropriate choice for investigating IIO with polytomous items. Furthermore, we used the following rules of thumb for the H^T coefficient: $H^T < .3$ implies that the item ordering is too inaccurate to be useful; $.3 \leq H^T < .4$ implies low accuracy of item ordering; $.3 \leq H^T < .4$ implies medium accuracy; and $H^T \geq .5$ implies high accuracy.

5.5 Results

Table 5.2 shows the results for the TEST, AISP, and IIO analyses with regard to the Emotional Stability and Conscientiousness scale.

5.5.1 Emotional Stability Scale

The AISP selected 11 of the 15 Emotional Stability items, resulting in a weak scale (H = .37) and six significant violations of IIO. Following the sequential procedure for method manifest IIO from Ligtvoet et al. (2010), item 4 (H_4 = .32) with three significant violations of IIO and the lowest H_i value of the items in the scale was removed. Reanalyzing the remaining 10 items, four items had one significant violation of IIO; item 5 (H_5 = .31), item 7 (H_7 = .45), item 12 (H_{12} = .41), and item 14 (H_{14} = .37). After removing item 5 (lowest H_i value), two items had two significant violations against IIO; item 7 (H_7 = .46) and item 12 (H_{12} = .42). After removing item 12, IIO holds for the remaining eight items with H^T = .57, indicating a high accuracy in the item ordering.

Table 5.2

Results of the Mokken analyses for the Emotional Stability and Conscientiousness scale.

	TEST					AISP					IIO				
Scale	*n*	# mono	*H*	# IIO	H^T	*n*	# mono	*H*	# IIO	H^T	*n*	# mono	*H*	# IIO	H^T
Emotional Stability	15	12(3)	.31	38(32)	.36	11	2(0)	.37	6(6)	.47	8	0(0)	.39	0(0)	.57
Conscientiousness	15	17(4)	.26	46(38)	.32	9	7(1)	.36	20(18)	.20	7	1(0)	.34	2(0)	.08

Note. *n* = number of items; # mono = number of violations of monotonicity; # IIO = number of violations of IIO.

To further investigate the item quality we plotted the rest-score functions. Rest-score functions for polytomous items are analogous to rest-score functions for dichotomous items, but now we use the mean item step response function. This mean response function should be monotonically increasing. For polytomous data, the summary functions of the item *step* response functions are needed to check whether polytomous *items* do not intersect (i.e., whether IIO holds). Sijtsma and Hemker (1998) showed that the sum of the item step response functions, written as the conditional expectation of the item score, can be used.

Inspecting the rest-score functions for the items for which IIO holds suggests that the IRFs are relatively far apart resulting in the reported H^T =.57. Thus for this scale, items can be ordered ranging from the most popular item "little faith in the future" (item 8) through the less popular item "needs confirmation from others" (item 3). Table 5.3 displays the Emotional Stability items and their item means ordered from most popular through less popular.

Table 5.3
Item content and item means of the Emotional Stability items for which IIO holds, ordered from most popular through less popular.

Item	item content	item mean
EMS8	much faith in future	3.56 (0.71)
EMS2	faith in own ability to tackle problems	3.30 (0.69)
EMS14	Faith in the use of personal contribution	3.25 (0.86)
EMS7	faith about own skills	2.84 (1.10)
EMS1	Not afraid of making mistakes	2.72 (1.12)
EMS10	rebound easily when things go wrong	2.46 (1.06)
EMS11	brooding for a long time over what went wrong (R)	1.84 (1.07)
EMS3	needs confirmation from others (R)	1.74 (1.04)

Note. EMS = Emotional Stability; (R) = reversed scored item, standard deviations are displayed between brackets.

5.5.2 Conscientiousness Scale

For the Conscientiousness scale we found that 9 out of 15 items were selected by AISP, resulting in a weak scale (H = .36), one significant but no serious violation of monotonicity for item 11 (H_{11} = .32, *Crit* = 30) and 18 significant violations of IIO. Removing item 5 (H_5 = .46 and seven significant violations of IIO) resulted in IIO

with H^T = .08, which implies that the item ordering is too inaccurate to be useful because the IRFs of the remaining items are close together. These results also suggest that this scale is measuring a very narrow construct. Inspecting the item content all items refer to working orderly and accurately.

5.6 Discussion

When a researcher investigates the quality of a scale he or she should always be aware of the fact that the intensity of the items is not automatically reflected in the ordering of the item means. Investigating IIO may tell a researcher that several items in the item pool have similar item response functions or item step response functions that do not allow for an ordering of items to severity. In fact, many personality scales are constructed like this. They often consist of a repetition of similar statements constructed around the pivotal question that asks for example if someone is organized. Related to the trend of online computer-based testing and the demand for short questionnaires, this information might be useful to shorten a questionnaire.

On the other hand, removing items from a scale that violate the assumption of IIO requires a delicate balance between different psychometric and content arguments, which in many psychometrically-oriented papers are not given much thought. When first selecting items with the AISP, we obtain high quality items with respect to their discriminating power. Items with relatively flat IRFs are eliminated. This implies that when removing items due to violations of IIO, we remove items that, in principle, are suited to scale persons according to their latent trait. A researcher then should weigh carefully whether the removal of items to obtain IIO, deteriorate the content validity of the scale and the reliability of its measurement. We want to stress that *item content* and its theoretical relevance should be studied carefully as a criterion in itself for evaluating and if necessary eliminating of items. To obtain a set of items that allow IIO, the final item set might measure the same concept through repeating similar questions, thus reducing the bandwidth of the concept being assessed. This already holds for selecting items with relatively high H_i values (Egberink & Meijer, 2011) and the bandwidth may be further reduced by removing items that violate the assumption of IIO. Therefore, a recurrent theme in the psychological literature is the necessity of *not* relying solely on the output of statistically defined or other analytical or mechanical procedures, in spite of all their possible sophistication. Removing items 4 and 5 from the Emotional Stability of the RPBF can be done without consequences for the psychometric quality of the scale,

because both H_i value were relatively low (H_4 = .32 and H_5 = .31) and the item content was also covered by other items in the scale.

As Meijer and Egberink (in press) discussed another observation is that for some clinical scales the H^T values can be low reflecting the fact that respondents find it difficult to distinguish one item from another with respect to intensity (e.g., PAR scale of the BSI). For other scales they found that groups of items were close together, with sometimes one or two items further away from these items (DEP and HOS scale of the BSI). These "outliers" were responsible for the high H^T coefficients.

In conclusion: when applied researchers use personality questionnaires they should realize that items or elements of trait characteristics may be differently ordered for persons with different sum scores or latent trait values, that different methods are available to investigate IIO and that investigating IIO also provides interesting diagnostic information about the general quality of a questionnaire.

Chapter 6
Unproctored Online Cognitive Ability Testing and Detecting Cheating

6.1 Introduction

Many organizations now publish their job vacancies on the Internet because this is a cheap and efficient way of advertising. If large numbers of applicants are involved, it is worthwhile automating parts of the selection process. As discussed in Chapter 1, various companies offer systems that automate much of the administrative process. Often these applicant tracking systems are also capable of delivering psychological tests and questionnaires. There are also companies, such as employment agencies and recruitment and selection agencies, that operate as brokers, matching job seekers to vacancies. They build up databases of large numbers of candidates. It is the quantity of relevant information available on these candidates that determines the value of the databases. Candidates who sign up are asked to provide information about themselves. They may also be requested, and sometimes required, to complete tests and/or questionnaires which can be done in an unproctored setting, usually at home. The organizations store the test data in their databases and use them to match candidates with vacancies. This has two major ramifications for the use of tests and questionnaires: the first is that candidates' test and questionnaire results can be used to determine their suitability for several jobs with different job descriptions, and the second is that questionnaires are completed in an unproctored setting. Therefore, in the present chapter, we investigate test-retest scores of an unproctored and proctored version of a cognitive ability test, using different psychometric methods.

I thank Jorge Tendeiro and Annette Maij-de Meij for their contribution to this chapter.
This chapter has been submitted for publication

6.2 Cheating and Detection of Cheating

A major disadvantage of unproctored online testing is that there can be no guarantee that candidates have taken the test themselves (Guo & Drasgow, 2010). Someone else may have taken the test for the candidate or may have helped the candidate during the test. This is a particular risk for cognitive ability tests, which demand a certain score before candidates can be admitted to the selection procedure. Because cognitive ability tests have a high predictive value for later job performance (Schmidt & Hunter, 1998), they are frequently used in the first selection step. Candidates whose cognitive ability level is too low may be excluded from the next step in the selection process. They will, therefore, do their best to obtain a high score on the test so that they can be considered for the job they are after. The risk here is that they will attempt to improve their score by cheating. Cheating is not only a problem with unproctored online testing. It occurs with various types of test, with different professional groups, and under different test conditions, as well as with proctored tests (Cizek, 1999, p. 73).

For the detection of cheating, we can distinguish between technological and statistical methods (Lievens & de Soete, 2011). Technological methods include supervision by webcam or biometrical identification tools such as key-stroke analysis (Foster, 2009). Statistical methods are used to analyze a candidate's response pattern (Bartram, 2008; Foster, Maynes & Hunt, 2008) or to verify a test score obtained on an unproctored test by means of a second, proctored, test (Guo & Drasgow, 2010; Makransky & Glas, 2011; Nye, Do, Drasgow, & Fine, 2008). The analysis of response patterns can focus on the time taken to answer each item or test, or on the relation between the item responses and the total test score. Fast or slow responses to items or unexpectedly correct or incorrect answers given the total score may indicate that the candidate had an answer key, had access to the test content, or received help from a third party (Meijer & Sijtsma, 2001). Using a second test to verify the score of an unproctored test is helpful to obtain information about the extent to which the first test score can be regarded as a realistic score for that candidate. Both technological and psychometric identification methods can point to possible cheating behaviour on the part of candidates (Tippins et al., 2006). However, it should be emphasized that these methods provide an indication only. Score discrepancies could also arise, for example, because a candidate feels pressure to perform well under a proctored condition.

The International Test Commission's (2005) guidelines for computer-based and online testing advise that an unproctored test administration should be followed by a second proctored test administration. In HRM research practice, there are few

studies available on cheating on unproctored cognitive ability tests. Whereas Oud, Bloemers, and Reitz (2009) have shown that it is possible to cheat in an unproctored online cognitive ability test, this does not mean that cheating is common practice. Nye et al. (2008) detected no cheating among a group of applicants in a test of perceptual speed. They found that only 0.5% of the applicants scored lower than 1.96 SD of the cut-off score on the verification test.

Lievens and Burke (2011) found similar results, with percentages between 0.3% and 2.2%. However, higher percentages are presented at international conferences on the basis of practical experience. Gibby (2010) reported that 9.6% of the applicants in an international selection program showed a large discrepancy between the score on the test taken at home and the verification test. Burke (2010) reports aberrant scores ranging from 8% through 11.7% in a verification test for various job groups. The highest discrepancy was found among recent graduates. Burke's (2010) explanation is that new graduates often know one another and apply for jobs at the same organizations at the same time.

Do, Shepherd, and Drasgow (2005) compared test-retest scores for various tests, including problem-solving ability tests, and found similar results for proctored and unproctored tests. Whenever differences were found in the mean scores of unproctored and proctored tests, the scores were often higher for the proctored tests (Lievens & Burke, 2011; Nye et al., 2008). Templer and Lange (2008) used a combination of proctored and unproctored delivery for a personality questionnaire and a cognitive ability test under time pressure. They found an increase in test scores between the first and the second test, but in their study this could also be attributed to a practice effect and not to the presence of supervision.

In the present chapter, we investigate test-retest scores of a proctored and unproctored version of a computerized adaptive test (CAT) for cognitive ability. Two recently proposed methods are applied and compared with respect to their effectiveness. Furthermore, we use diagnostic information obtained from studying individual item score patterns to interpret unexpected test-retest results. To our knowledge, for IRT-based adaptive testing only simulation studies have been carried out with regard to score differences between tests taken at home and verification tests, and, as discussed above, earlier studies on fixed length test-retests did lead to ambiguously interpretable results. Before we discuss an empirical study, we first provide an overview of proctored and unproctored testing using computer-based administration methods.

6.3 Cheating and the Validity and the Utility of Selection Procedures

Organizations are increasingly using unproctored online testing. Consequently, there is a shift in the research question – from "Is unproctored online testing suitable for cognitive abilities?" to "How can this testing method be used without severely reducing the reliability and validity of the selection procedure?" (Tippins, 2009a). If dishonest behaviour reduces a test's validity, this will also reduce the test's utility for the unproctored online testing of cognitive ability for the purpose of selection decisions. Weiner, Knapp, and Hogan (2011) showed how cheating can affect the validity of a test. Using simulations in which they varied the percentage of cheaters and the difference between the score obtained through cheating (short: cheating score) and the true score, they estimated the validity and the expected decision errors (type I errors). They concluded that, in general, cheating has a negligible impact on the validity and on the selection decisions if the percentage of cheaters is lower than 10%, the difference between the cheating scores and the true scores is less than 1 SD, and a low to average score is used as the cut-off score. Under extreme circumstances, however, cheating can have a dramatic impact. If there is a high percentage of cheaters and if the differences between the cheating score and true score are larger than 1 SD, this can considerably reduce the validity and the utility and can increase the number of selection errors to 40%.

Weiner (2010) has also studied the influence of dishonest behaviour on test utility using Brogden's (1949) formula. Assuming a starting salary of €25,000 and a difference between top and average performance of 40%, this would yield a cost of €18,000 per 100 candidates in the event of 5% cheats and a difference of 2 SD between the cheating score and the true score. With 5% cheaters and a difference between the cheating score and the true score of 1 SD, the costs would be €10,000 per 100 candidates. With these calculations in mind, it can be concluded that the more effectively cheating is detected, the more an organization will benefit.

6.4 Cognitive Ability Tests and Verification Tests in the Selection Process

In practice, organizations differ in their considerations regarding the use of cognitive ability tests and how they use test outcomes in the selection process. Some organizations prefer to meet the candidates personally and to use cognitive ability

tests only during a proctored assessment. Others use unproctored testing before the actual assessment (pre-selection): candidates who do not meet the required intelligence level for a job are not invited for an interview. These organizations view the reduction in personnel costs and the cost of hiring test venues as a distinct advantage.

As outlined above, unproctored online testing carries the risk that someone other than the candidate has taken the test or that the candidate received help during the test. This may imply that candidates who do not meet the required cognitive ability level may be invited for a job interview. Organizations are aware of this risk and deal with it in different ways within their selection procedures (Tippins et al., 2006). Some organizations accept the risk and do not use a verification test. They assume that candidates who lack the requisite intelligence will not reach the next selection round. For example, they expect unqualified applicants to fail in job interviews with various managers from the organization. Other organizations are less strict in their application of the cut-off score for the cognitive ability test. They base their final judgment about job suitability on a combination of factors, such as IQ score, personality profile, role-play outcomes, interviews, and curriculum vitae information. Thus, a lower score on the cognitive ability test may be compensated by a favourable personality profile.

Organizations that find the detection of cheating in the unproctored test important require all candidates to take a proctored verification test following the pre-selection stage. This is true for banks, for example, where integrity is an important cultural value (Tippins et al., 2006; Weiner et al., 2011). These organizations cannot afford to hire on staff who have behaved dishonestly during the selection process. Research shows that warnings about checking test scores can have a deterrent effect on cheating (Lerner & Tetlock, 1999), which is why some organizations use the verification test primarily as a strategy to discourage such behaviour. Before candidates take the unproctored test, they are told that the test results may be checked. A random sample of candidates who have scored above the cut-off point on the unproctored test are then selected to take the verification test.

The use of an unproctored online cognitive ability test holds a particular economic appeal if large numbers of new employees have to be recruited each year. This applies, for example, to the regular recruitment of trainees by large organizations. There may be a greater risk of dishonest test-taking with this group of candidates than with others, because many know one another from university and they tend to look for similar first jobs at the same time. Burke (2010) identifies 3-4% more cheaters in this group than for applicants with several years of work

experience. When choosing which procedure to use for the unproctored online test and verification test, an organization will also have to take into account the group the candidates are selected from and the probability that candidates will know one another.

6.5 Using Verification Tests to Detect Aberrant Scores

Three groups of statistical verification methods to detect aberrant scores can be distinguished in the literature. The first method uses person-fit statistics (e.g. Meijer & Sijtsma, 2001). The response pattern of the proctored test is studied and compared with the test score and the response pattern of the unproctored test (e.g., Tendeiro, Meijer, Schakel, & Maij-de Meij, in press). When using this method, inconsistencies in a candidate's response pattern on the proctored test is seen as an indication of cheating. The second group of methods uses a sequential verification test to explore the extent to which the score on the unproctored test can be confirmed (e.g., Makransky & Glas, 2011). In a sequential verification test, items are presented one at a time. After each response, a decision algorithm is used to evaluate whether the score on the earlier test is consistent or aberrant. The test ends as soon as decision can be made with a sufficient degree of certainty, which means that sequential verification tests do not have a fixed length. A simulation study by Makransky and Glas (2011) showed that this type of verification test provides the same detection power at a quarter the length of a verification test containing a fixed set of items. The test result indicates whether the score on the unproctored test should be accepted or rejected. When a score is rejected, the test does not provide an alternative score.

In this chapter we focus on a third group of methods that compare the scores on two separate tests. Nye et al. (2008) used a method in which the standardized score on the selection test is corrected for the regression effect towards the mean. The corrected score was compared with the standardized score on the verification test using paired-sample t-test scores. Lievens and Burke (2011) also applied this method in a study of score differences between proctored and unproctored cognitive ability tests for a group of job applicants.

Guo and Drasgow (2010) conducted a simulation study to compare two detection methods that were designed to identify cheating in IRT-based CAT: a z-test and a likelihood ratio test (LRT). Their study showed that the z-test had a higher or similar power than the LRT in most cases. They also showed that longer

tests were better able to detect aberrant scores in both proctored and unproctored settings.

The verification test can also be used in another way (Weiner et al., 2011). A simple decision rule can be used that demands that participants in both the unproctored selection test and the verification test must score above the cut-off point. An advantage is that this method can be clearly and simply explained to the candidate and that the verification test is at the same time also a selection test. However, this strategy does not compensate for differences that arise through inaccurate measurement. In the present study the effects of different detection methods on the percentage of detected aberrant test-retest scores in a practical test situation are compared.

6.6 Method

6.6.1 Instrument

Cognitive Ability Test

An online IRT-based CAT for cognitive ability was used. There are two versions of the test: the Connector Ability (Maij-de Meij, Schakel, Smid, Verstappen, & Jaganjac, 2008), which can be delivered online in an unproctored setting, and the Connector Ability Validator, a verification test which can only be used on location in a proctored setting.

The Connector Ability measures the general cognitive ability level by means of three subtests: Figure Series, Matrices, and Number Series. The Connector Ability aims at educational levels ranging from secondary educational level to university degree and has three norm groups (secondary educational level, higher educational level, and university degree). The test is primarily used for selection. A key design principle underpinning the Connector Ability test is that candidates' cultural backgrounds should have minimal influence on the test score. Words are kept to a minimum in the test items and the principles that apply to the items are described at length in the instructions (see Figure 6.1, for an example).

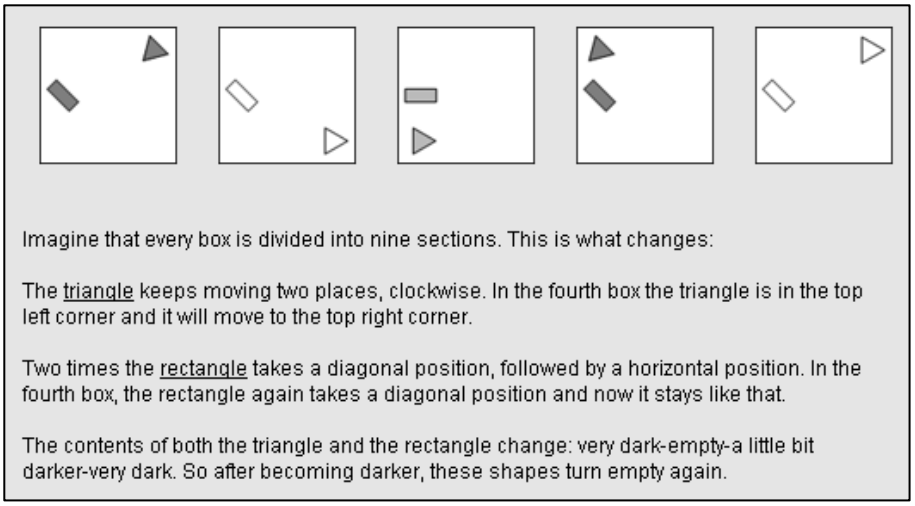


Figure 6.1: An example of the explanation of the principles underpinning the items in the instructions for the Figure Series subtest of the Connector Ability.

Practice items are used to check whether the candidate has understood the instructions. Prior to the actual assessment, the candidate can take a practice test at home, which contains the same type of items and instructions. The practice test is not adaptive and has a fixed set of 14 items per subtest. When candidates finish the practice test, they immediately receive a personal report by e-mail, informing them of their general intelligence ('*G* factor') score. They are not given any feedback regarding which items they answered correctly or incorrectly.

The test developers gained experience with this test in 2008 and 2009, and they expanded the item bank by including experimental items during delivery. In 2010, two versions of the test were developed on the basis of the available item bank – the Connector Ability and the Connector Ability Validator (hereafter referred to as Validator; Maij-de Meij & Schakel, 2011). The Connector Ability can be used in both unproctored and proctored selection settings. The Connector Ability has an item bank of several hundred items for each subtest. A minimum of 10 items and a maximum of 15 items are presented per subtest. The sequence for the alternative answers is randomized with each test delivery. The Validator was specially developed as a test to verify a candidate's score on the unproctored test. A fixed-length set of 7 items per subtest is presented to the candidate. The Validator has an item bank of about 50 items per subtest. Items with a high discrimination parameter ($a \geq .80$) were selected in order to obtain a reliable measurement more quickly and with fewer items. The Validator does not provide any instructions beforehand, because candidates already received these when taking the Connector Ability. This means

that the test is about 30% to 50% shorter than the Connector Ability and can be completed on average within 30 minutes. The Validator is only administered in a proctored setting in order to prevent candidates from becoming familiar with and telling others about the test items. Organizations must confirm to administer the test only in a proctored setting.

Information for Candidates

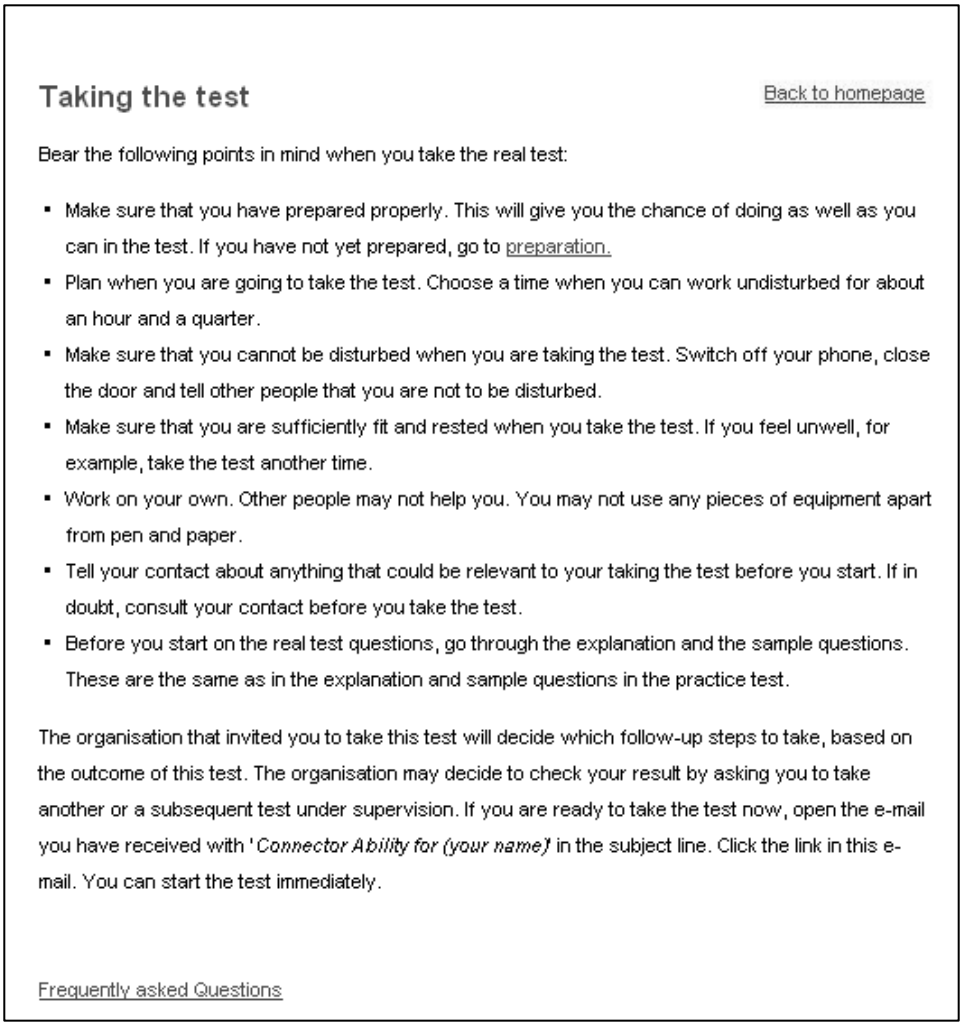
From the candidates' point of view it is important to obtain a high test score, as this increases their chances of being admitted to the next step in the selection process. Candidates can apply for a practice test before the actual test as often as they want, but the same practice test is administered each time. The fact that two-thirds of all candidates preparing for the Connector Ability complete the practice test two or more times shows that candidates like to be well prepared.

Once candidates have been invited to take the unproctored online Connector Ability, they have one week to take the test. They receive a link to a website providing information about the practice test, the possibility of a verification test, minimum requirements for their computer, advice on optimum test conditions, personal preparation, and telephone support (see Figure 6.2, for a screenshot of the website). Candidates are expected to prepare well for the test and to make sure that they take it under optimum test conditions.

When the candidate has finished the Connector Ability, the test system can send the results to the assessor only, to the candidate only, or to both. To date, all the organizations that have used the Connector Ability have opted to send the results only to the assessor, as they themselves prefer to inform the candidate, either orally or in writing, of the result. Until now, no organization has chosen to send a candidate an automatic rejection.

6.6.2 Detecting Aberrant Scores and Taking Decisions

When administering the Connector Ability followed by verification with the Validator, the assessor is faced with two test scores. These test scores are seldom identical, due to measurement error. The assessor must decide, on the basis of these two different scores, whether or not the candidate has completed the unproctored test honestly and may proceed with the selection process. Sometimes there are large differences between the two scores, with the candidate insisting that he or she did not cheat on the unproctored test. While this is statistically possible, it only applies to a small percentage of cases. This procedure confronts the assessor with a problematic decision. While a decision rule can be of help here, in practice assessors



Taking the test

Back to homepage

Bear the following points in mind when you take the real test:

- Make sure that you have prepared properly. This will give you the chance of doing as well as you can in the test. If you have not yet prepared, go to preparation.
- Plan when you are going to take the test. Choose a time when you can work undisturbed for about an hour and a quarter.
- Make sure that you cannot be disturbed when you are taking the test. Switch off your phone, close the door and tell other people that you are not to be disturbed.
- Make sure that you are sufficiently fit and rested when you take the test. If you feel unwell, for example, take the test another time.
- Work on your own. Other people may not help you. You may not use any pieces of equipment apart from pen and paper.
- Tell your contact about anything that could be relevant to your taking the test before you start. If in doubt, consult your contact before you take the test.
- Before you start on the real test questions, go through the explanation and the sample questions. These are the same as in the explanation and sample questions in the practice test.

The organisation that invited you to take this test will decide which follow-up steps to take, based on the outcome of this test. The organisation may decide to check your result by asking you to take another or a subsequent test under supervision. If you are ready to take the test now, open the e-mail you have received with '*Connector Ability for (your name)*' in the subject line. Click the link in this e-mail. You can start the test immediately.

Frequently asked Questions

Figure 6.2: Connector Ability: Screenshots of the website where candidates can find all the information they need about the test procedure before taking the test.

do not have a sufficient statistical background to interpret aberrant scores properly, to explain them, and to defend them to the candidate. An automated psychometric detection method with a clear report on the extent of the aberrant score in relation to the unproctored test can be useful in supporting an assessor in this decision.

In the Validator report the scores on the Connector Ability and the Validator are classified into two categories. The report displays a 'green light' when the scores for the unproctored (Connector Ability) and proctored (Validator) test are comparable or when the score for the proctored test is higher than the score for the

unproctored test. In the Validator report only scores for the unproctored test are reported. As a result, there is no confusion involving a second, slightly different, score. An 'amber light' is reported when the score on the Validator is significantly lower than the score on the Connector Ability (see Figure 6.3, for an example).

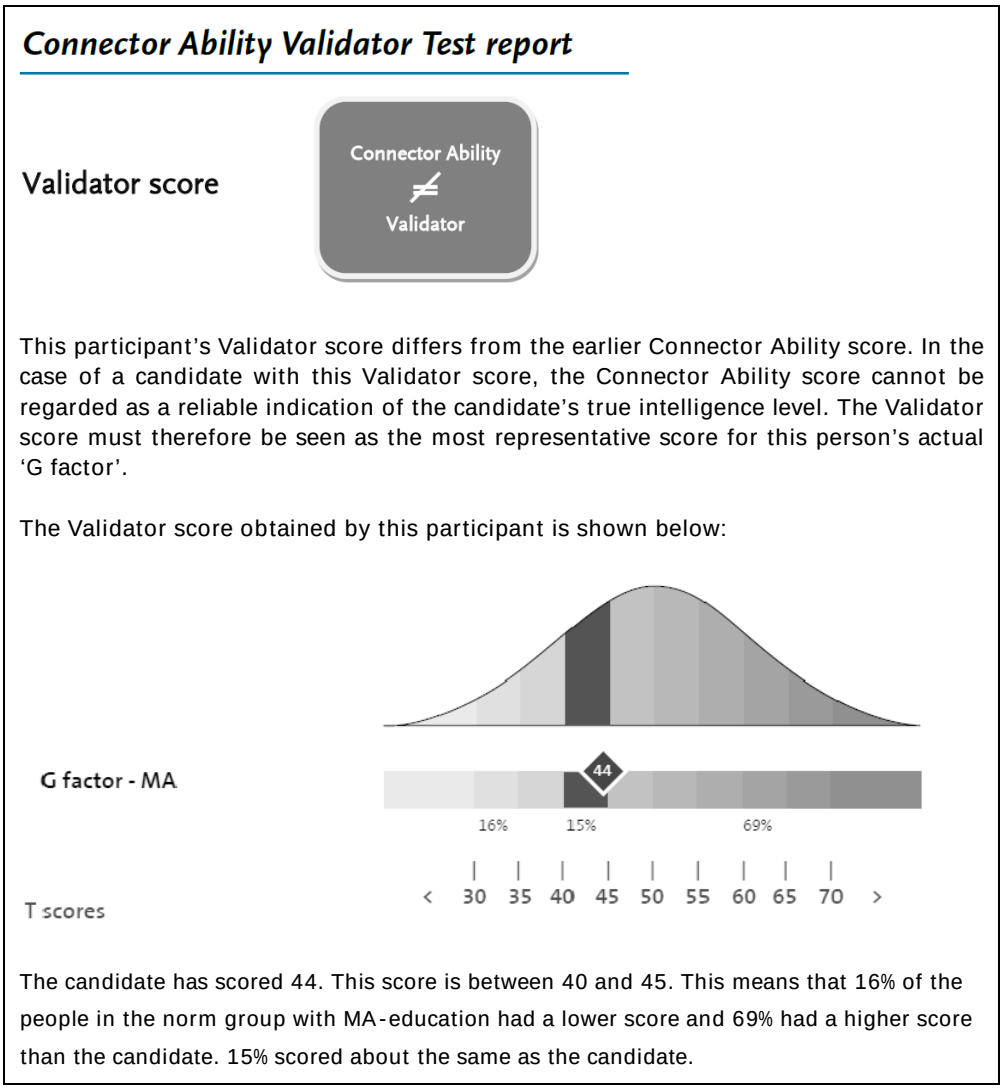


Figure 6.3: Validator report with an "amber light" when the score on the unproctored (Connector Ability) test differs much from the proctored (Validator) test.

The test developers decided to use an amber light rather than a red one, because although a significant aberrant score may indicate possible dishonest behaviour, it does not automatically rule out a candidate. When a candidate receives an amber light, the report will show the scores for the Validator instead of the score of the Connector Ability as in the green light report. The reliability interval for the Validator score is taken into account in the comparison of scores. If the upper boundary of the reliability interval is smaller than the *θ*-estimate ($\hat{\theta}$) on the Connector Ability, the scores are considered aberrant. For example, if $\hat{\theta}$ on the Connector Ability is .56 and $\hat{\theta}$ on the Validator is -.12, the maximum $\hat{\theta}$ on the Validator is calculated given a particular confidence interval. With a standard error of .42 and $\alpha = .05$, the maximum $\hat{\theta}$ on the Validator is .57 (see section 6.6.4 Analyses, for the formula). The upper boundary of the confidence interval of the Validator score is larger than the $\hat{\theta}$ on the Connector Ability. The Validator score is, therefore, not classified as aberrant.

6.6.3 Sample

Data were collected in the first half of 2011 from applicants for various jobs with different organizations. All candidates took the Connector Ability in an unproctored setting of their choice, usually at home. The Validator was then administered in a supervised test location at the organization. Thus, the sample only included participants who after completing the Connector Ability at home were selected for a next selection round. The mean time between taking the Connector Ability and the Validator was 11.49 days (*SD* = 10.34).

The sample comprised 425 persons, with a mean age of 27 years (*SD* = 9.7) with 69%, males, 28% females; for 3% gender was unspecified. 61% had no work experience, 58% had a university degree, 27% had higher education, and 15% had secondary education. 80% were Dutch natives, 10% Western immigrants, and 10% non-Western immigrants. 84% were applying for a job at university degree level, 14% at higher educational level and 2% at secondary educational level.

6.6.4 Analyses

We compared the outcomes of two different methods for detecting aberrant scores between the Connector Ability and the Validator; the z-test (Guo & Drasgow, 2010) and the Validator method (Maij-de Meij & Schakel, 2011).

Let $\hat{\theta}_{CA}$ denote the ability estimates from the Connector Ability and let $\hat{\theta}_V$ denote the ability estimates from the Validator test and denote their standard errors as SE_{CA} and SE_V, respectively. Under normal response behavior it is expected that $\theta_{CA} = \theta_V$; if a candidate obtains correct answers on the Connector Ability as a result of cheating, it is expected that $\theta_{CA} > \theta_V$. Thus, we test: $H_0 : \theta_{CA} = \theta_V$ versus $H_a : \theta_{CA} > \theta_V$. Using maximum likelihood estimation (MLE), each MLE is asymptotically normal, so their difference is also asymptotically normal. Given the assumption of conditional independence under H_0, $\hat{\theta}_{CA}$ and $\hat{\theta}_V$ will be independent. Therefore, under H_0, the standardized score difference between the two tests follows a standard normal distribution. A z-statistic can be computed as follows:

$$Z = \frac{\hat{\theta}_{CA} - \hat{\theta}_V}{\sqrt{SE_{CA}^2 + SE_V^2}} .$$

Depending on the desired α-level, z-test values reflect aberrant test scores. Thus $z_{1-\alpha}$ can serve as the z-value for the standard normal distribution ((1 - α)*100% confidence limit). In the study of Guo and Drasgow (2010), a one-tailed test with type I error α = .01 was conducted. So if the z-statistic was 2.33 (= z_{99}) or larger, the candidates were classified as having cheated in the unproctored test.

At the Company a slightly different method is being used: the Validator method. In the comparison of the scores on the Connector Ability and Validator, the maximum score on the Validator is defined as:

$$\hat{\theta}_{VU} = \hat{\theta}_V + z_{1-\alpha} * SE(\hat{\theta}_V)$$

where $\hat{\theta}_V$ is the estimated θ on the Validator, $z_{1-\alpha}$ is the z-value for the standard normal distribution ((1 - α)*100% confidence limit), $SE(\hat{\theta}_V)$ is the standard error of $\hat{\theta}_V$, and $\hat{\theta}_{VU}$ is the upper boundary of the confidence interval of $\hat{\theta}_V$. When $\hat{\theta}_{VU}$ is smaller than $\hat{\theta}_{CA}$, the score on the Validator ($\hat{\theta}_V$) is considered aberrant.

The methods differ in the principles they apply for establishing aberrant scores. The Validator method only takes the uncertainty of $\hat{\theta}_V$ into account, whereas the z-test takes both the uncertainty of $\hat{\theta}_{CA}$ and $\hat{\theta}_V$ into account. The two methods were compared in combination with three different type I errors (α=.05, .025 and .01 with respective z-values of z_{95} = 1.65, z_{975} = 1.96 and z_{99} = 2.33).

Table 6.1

Mean θ-values and standard deviations for the groups without and with aberrant scores for each detection method.

		no aberrant scores					aberrant scores				
			Connector Ability		Validator			Connector Ability		Validator	
method	ζ-value	*n*	mean	SD	mean	SD	*n*	mean	SD	mean	SD
ζ-test	1.65	391	0.73	0.63	0.86	0.62	34	1.05	0.68	-0.01	0.56
	1.96	401	0.74	0.62	0.85	0.62	24	1.05	0.75	-0.13	0.60
	2.33	406	0.74	0.62	0.84	0.62	19	1.04	0.82	-0.22	0.63
Validator method	1.65	361	0.71	0.60	0.90	0.60	64	1.04	0.74	0.20	0.65
	1.96	373	0.71	0.60	0.89	0.61	52	1.06	0.76	0.14	0.66
	2.33	385	0.72	0.62	0.87	0.62	40	1.12	0.67	0.09	0.60

6.7 Results

Table 6.1 presents the mean θ-values and their standard deviations for the group of persons classified as normal (no aberrant score) and aberrant for different cut-off scores. For the complete group of candidates, the mean score on the Connector Ability did not differ significantly from the scores on the Validator ($t(424) = -1.37$, $p = .17$, Cohen's $d = .06$; 95% CI [-.091, 0.162]). Candidates with normal score fluctuations scored higher on the Validator than on the Connector Ability (when using $z = 1.96$: $t(400) = -4.54$, $p < .00$, Cohen's $d = .17$; 95% CI [-.158, - .062]). This is consistent with earlier studies (Nye et al., 2008), which also reported higher scores on the verification test.

As can be verified from Table 6.1, the percentages of classified aberrant scores are 8.0%, 5.2%, and 4.5%, respectively for the z-test, and 15.1%, 12.1%, and 9.4% for the Validator method, respectively for $\alpha = .05$, $\alpha = .025$ and $\alpha = .01$. The results show that the Validator method classifies more scores as aberrant than the z-test. Also, as expected, the use of a larger confidence interval results in a lower percentage of aberrant scores.

Figure 6.4 shows scatterplots of the scores on the two tests for both methods with $z = 1.96$ ($\alpha = .025$) for both candidates classified as aberrant and normal.

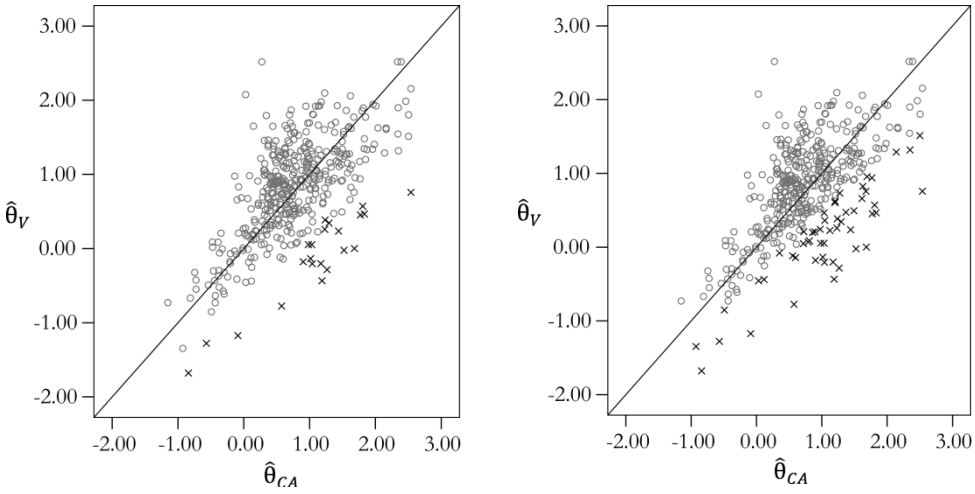


Figure 6.4: Scatterplots of the scores on both tests. Scores classified as aberrant, using a z-value of 1.96 ($\alpha = .025$), are indicated by 'x'. The left panel shows the results for the z-test and the right panel for the Validator method.

Candidates with scores below the diagonal scored lower on the Validator than on the Connector Ability. It is clear that the methods differ in the degree to which they consider some scores below the diagonal to be aberrant. The z-test accepts a larger distance below the diagonal than the Validator method and considers therefore less scores as aberrant. To obtain more diagnostic information about the test scores for persons classified as normal and aberrant, respectively, we investigated the configuration of item scores for these different groups.

6.7.1 Diagnostic information using the CUSUM

Cheating may result in unexpected item score patterns. When a person gets help from another more able person on a subset of items and as a result answers many items correctly, strings of correct scores are observed. These strings of correct scores are unexpected given the candidate's trait value. For example, Jacob and Levitt (2003) provided empirical evidence that strings of correct answers on an educational test were due to cheating. In their study, teachers changed incorrect item scores into correct scores to raise students' total scores so that the school was evaluated more positively. Our testing context is different, but the cheating mechanism is similar: a person's test score may be raised through the help of another more able person. Therefore, to detect these types of unexpected item score patterns, we used the following strategy in addition to calculating the z-scores.

We determined the likelihood of an item score pattern on the Validator (i.e., *proctored* test) using the estimated latent trait value on the Connector Ability (i.e., *unproctored* test) through a Cumulative Sum procedure (CUSUM, Meijer & van Krimpen-Stoop, 2010). The CUSUM procedure can be considered as a person-fit procedure that is sensitive to strings of unexpected item scores given the estimated latent trait value. Bradlow, Weiss, and Cho (1998) and van Krimpen-Stoop and Meijer (2000, 2001) proposed statistics that are based on a CUSUM procedure. Like other person-fit statistics, a researcher can specify a type I error (or control limit) on the basis of which an item score is classified as normal or aberrant. For each item, i, in the test, a statistic, T_i, can be calculated that is a weighted version of the residual X_i - $P_i(\theta)$, where $P_i(\theta)$ is the probability of giving a correct answer to item i calculated using a specific IRT model and X_i is the observed item score (in the present study, '0' for an incorrect answer and '1' for a correct answer). In this study the mean residual was used, that is, $[X_i - P_i(\theta)] / k$, where k is the number of items in the CAT. Then, the sum of these T_is equals:

$$C_i^+ = \max[0, T_i + C_{i-1}^+],$$
$$C_i^- = \min[0, T_i + C_{i-1}^-],$$

and

$C_0^+ = C_0^- = 0.$

Thus, C^+ and C^- reflect the sum of consecutive positive and negative residuals, respectively. Let UB and LB denote appropriate upper and lower bounds. Then when C^+ > UB or C^- < LB, the item score pattern is classified as normal. For further technical details see van Krimpen-Stoop and Meijer (2001) and for recent developments see Armstrong and Shi (2009), and Tendeiro and Meijer (in press). In this study, we used the C^- CUSUM procedure because we were only interested in unexpected strings of '0' scores because this may reflect unexpected low scores on the Validator.

Table 6.2

$\hat{\theta}_{CA}$ and $\hat{\theta}_V$, together with the related z-values for the selected 20 persons.

	Connector Ability		Validator		
Person	$\hat{\theta}_{CA}$	SE_{CA}	$\hat{\theta}_V$	SE_V	z-value
449	2.33	0.37	0.15	0.25	4.87
516	2.01	0.34	-0.13	0.23	5.20
544	1.47	0.30	-0.12	0.21	4.36
577	1.03	0.29	-0.77	0.18	4.31
590	1.91	0.38	-0.31	0.28	4.73
674	3.84	0.60	0.93	0.26	4.46
694	2.66	0.45	0.06	0.35	4.56
752	1.87	0.35	-0.06	0.26	4.39
797	0.13	0.22	-1.23	0.18	4.87
848	1.67	0.32	-0.02	0.23	4.31
461	1.26	0.30	1.23	0.38	0.06
500	0.75	0.26	0.74	0.25	0.01
526	1.13	0.36	1.12	0.28	0.02
527	0.06	0.24	0.05	0.20	0.03
574	1.06	0.32	1.05	0.30	0.04
686	1.71	0.33	1.69	0.44	0.03
788	1.09	0.29	1.06	0.32	0.07
806	1.81	0.33	1.81	0.38	0.01
814	0.81	0.28	0.78	0.26	0.06
816	-0.07	0.20	-0.06	0.22	0.00

Note that we did not use $\hat{\theta}_V$ because a well-known problem is that when there is intensive cheating, this cheating results in a very consistent response pattern (almost all '1' scores) given the high trait score. As a result, there is no difference between a high-ability examinee giving many correct answers as expected and a low ability examinee cheating on the test. Thus, when both tests are answered by the same person under similar conditions, the item score pattern on the proctored Validator will be classified as normal. However, when a candidate cheats on the unproctored Connector Ability, and as a result obtains a high trait score, this trait score is unexpected given the configuration of the items scores on the Validator.
To illustrate the CUSUM procedure as an additional diagnostic tool to the z-test, we selected 10 persons with the largest differences between the unproctored en proctored scores (largest z-values). Furthermore, we selected 10 persons with almost similar scores on the Connector Ability and the Validator. In Table 6.2 we depicted $\hat{\theta}_{CA}$ and $\hat{\theta}_V$, together with the related z-values for these 20 persons. In Figure 6.5 the CUSUM charts are presented for a number of interesting cases. Abilities were estimated using maximum likelihood estimation (MLE) for the 2-parameter-logistic model. Control limits were estimated for each CUSUM statistic and for each candidate. For each sample, we computed bootstrap distributions for the 1% and 5% control limits (number of resamples equal to 1000). Our estimates were computed as the medians of the corresponding bootstrap distributions. The median was used because we observed that the bootstrap distributions were often nonsymmetric and/or multimodal.

In the left panels of Figure 6.5, CUSUM charts are shown for persons that were classified as aberrant using the z-test and on the right panel CUSUM charts are shown for persons that were classified as normal by the z-tests. For the aberrant z-test cases, for 5 cases the CUSUM crossed the 5% control limit, and for 4 cases the 1% control limit (not tabulated). More interesting is, however, that these charts inform the assessor which items in the test are answered according to the IRT model, and which items are answered in an unexpected way. Consider Person 797 (aberrant) and Person 527 (normal). Both persons have a $\hat{\theta}_{CA}$ of around 0 on the Connector Ability, yet, their response behavior on the Validator is different. Person 527 answers the items on the Validator as expected: the item scores and the expected scores on the basis of the IRT model are similar resulting in a rather flat CUSUM chart. Person 797, however, has many large differences between observed and expected scores, resulting in a decreasing chart. In Table 6.3 we depicted the

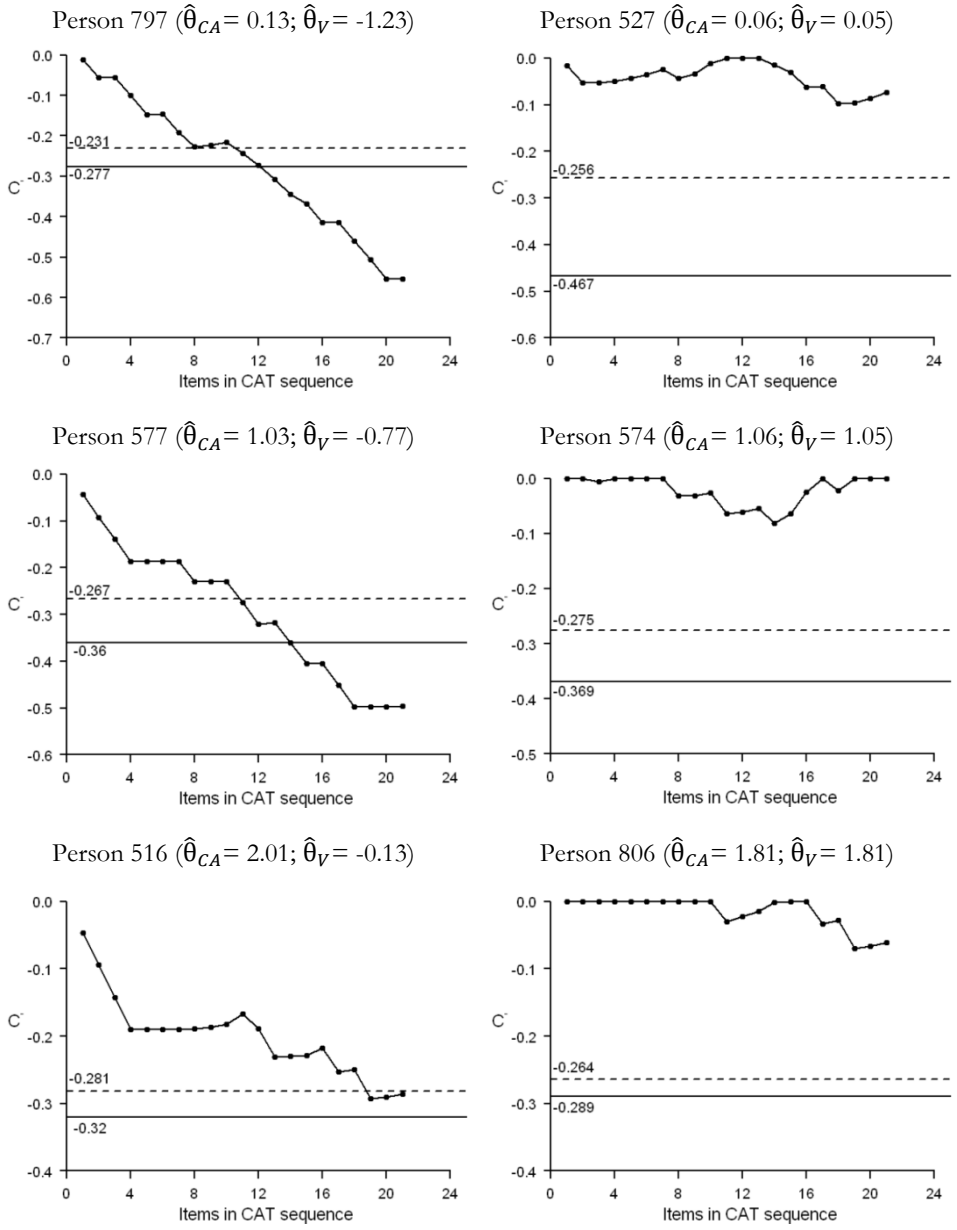


Figure 6.5: CUSUM charts for 6 different cases. *Note:* the horizontal black line indicates the 5% control limit and the horizontal dashed line the 1% control limit.

CUSUM procedure for this person. Note that only 6 items of the 21 items are answered correctly, and that many items should be relatively easy for this person. Note that Items 11 through 16 (thus 6 items in a row in this CAT) are answered incorrectly, which is very unexpected given the person's trait value, $\hat{\theta}_{CA} = 0.13$, the item parameters, and also given the adaptive nature of a CAT.

Persons 577 and Person 574 both have a $\hat{\theta}_{CA}$ of around 1, but, again, Person 577 answers many items incorrectly. For Person 516 (aberrant) it is interesting that he/she starts with three incorrect answers that are very unexpected (large drop in the chart), but then answers 7 items in a row correctly. Four of these items are measuring Figure series and three items are measuring Matrices. Thus, this answer pattern may be related to the item content.

Table 6.3
CUSUM procedure for Person 797

Item	Score	$P_i(\hat{\theta})$	T_i	C^-
1	0	0.247	-0.012	-0.012
2	0	0.909	-0.043	-0.055
3	1	0.996	0.000	-0.055
4	0	0.947	-0.045	-0.100
5	0	0.980	-0.047	-0.147
6	1	0.986	0.001	-0.146
7	0	0.969	-0.046	-0.192
8	0	0.736	-0.035	-0.227
9	1	0.920	0.004	-0.223
10	1	0.820	0.009	-0.215
11	0	0.604	-0.029	-0.244
12	0	0.625	-0.030	-0.273
13	0	0.719	-0.034	-0.308
14	0	0.765	-0.036	-0.344
15	0	0.510	-0.024	-0.368
16	0	0.964	-0.046	-0.414
17	1	0.998	0.000	-0.414
18	0	0.960	-0.046	-0.460
19	0	0.979	-0.047	-0.506
20	0	0.999	-0.048	-0.554
21	1	1.000	0.000	-0.554

Note. $P_i(\theta)$ = probability of endorsing an item given $\hat{\theta}$; T_i = difference between the observed and expected score; C^- = minimum value of the CUSUM.

6.8 Conclusions and Discussion

The results showed that the χ-test is more conservative than the Validator test. As expected, the chosen type I error affected the number of candidates who were detected as cheaters. Nye et al. (2008) found only four aberrant scores in 856 tests (that is .5%). Lievens and Burke (2011) also used Nye et al.'s method and reported 1.0% through 1.8% for a numerical test. These percentages are lower than the percentages found in the present study. However, these studies did not use a CAT. In a CAT, $\hat{\theta}$ and the corresponding *SE* are available for each candidate, which may lead to more accurate measurement for each individual.

In real testing situations, it is impossible to know what percentage of applicants exhibited *true* dishonest and misleading behaviour when taking the test. The percentages found in the current study, 5% with the χ-test and 12% with the Validator method, for $\alpha = .025$ are consistent with the 8.1% to 11.7% range reported by Burke (2010) and the 9.6% found by Gibby (2010). Weiner (2010) and Weiner and Ruch (2006) have shown that percentage cheaters between 5% and 10% have only a limited impact on the validity and utility of the test. Thus, the percentage of aberrant scores found in the present study is not of such a magnitude that it threatens the use of the Connector Ability for unproctored online testing.

In the present study, the extent of cheating was based on differences between the scores on an unproctored test and the scores on a verification test. For diagnostic purposes we used the CUSUM method to further study the configuration of test scores. In a personnel selection context this may be an interesting way of helping psychologists and other assessors to obtain a picture of the candidate's response behaviour. In computer-based testing often test scores and candidate's reports are generated automatically. Besides information about test scores and subtest profiles, a CUSUM chart may be added that gives diagnostic information about irregular response behaviour. Future research into response patterns may shed light on whether this kind of indicators provides useful additional information over and above a comparison of total scores.

An alternative to the CUSUM strategy used in this study consists of estimating, for each candidate, two posterior distributions of ability using data from the unproctored and proctored tests. Both posteriors are then compared using the Kullback-Leibler divergence (KLD; see Belov & Armstrong, 2010; Belov, Pashley, Lewis, & Armstrong, 2007). Large values of the KLD indicate a significant change in performance between both tests. Critical values for the KLD at fixed levels of significance can be estimated using either simulation, approximating distributions such as the lognormal (Belov & Armstrong, 2010), or theoretical distributions which

the KLD follows under specific conditions (Belov & Armstrong, 2011). As observed by an anonymous reviewer, the KLD approach takes into account all available information from the posterior distributions, unlike other statistics which rely only on the first moments of the posteriors (e.g., Guo and Drasgow's z statistic). We observe that the CUSUM technique is of a different nature than the KLD. CUSUMs are sequential procedures which take into account the order in which the items are presented to each candidate. The KLD technique estimates posteriors from two sets of items (in our setting, the unproctored and proctored tests), but the order of the items within each set is not taken into account. In particular, psychometric information in the shape of CUSUM charts is not readily available for the KLD. Thus, the CUSUM and the KLD approaches can be regarded as two alternative ways for detecting aberrant response behavior. CUSUMs are specially suitable for situations where it is important to take into account a specific ordering of the items (e.g., the administration order).

The sample in this study consisted of persons with above-average university degree level. Consequently, for many candidates the test was relatively easy. One option in such a situation may be not to view at aberrant verification test scores above a certain value (e.g., .5 SD above the norm group mean). This will preclude less relevant aberrant scores in the right-hand side of the distribution, caused by inaccurate measurements in that area.

When deciding on a decision rule, it is also advisable to keep in mind the effects on applicant expectations (Schreurs, Derous, Proost, Notelaers, & De Witte, 2010). If a larger type I error is chosen, more candidates will be classified as aberrant on the basis of the verification test. Some of these candidates will be incorrectly labelled as cheaters. The possible negative impact of this on candidate expectations can reflect poorly on a organization's image. There is clearly a need for research into the effects of using a verification test after an unproctored online test.

References

Almagor, M., Tellegen, A., & Waller, N. G. (1995). The big seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of Personality and Social Psychology, 69*, 300-307.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika,* 43, 561-573.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement,* 36, 277-300.

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33*, 391-410.

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1-16.

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19*, 289-303.

Baron, H., & Austin, J. (2000). *Measuring abilities via the internet: Opportunities and issues*. Paper presented at the annual conference of the Society of Industrial and Organizational Psychology. New Orleans, LA.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*, 9-30.

Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *International Journal of Selection and Assessment, 8,* 261-274.

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185-1203.

Bartram, D. (2006). The internationalization of testing and new models of test delivery on the Internet. *International Journal of Testing, 6*, 121-131.

Bartram, D. (2008). The advantages and disadvantages of on-line testing. In S. Cartwright, & C. L. Cooper (Eds.), *The Oxford handbook of personnel psychology.* (pp. 234-260). Oxford: Oxford University Press.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement, 34*, 379-392.

Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback-Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology, 64*, 291-309.

Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback-Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7-14). Tokyo: Universal Academy Press.

Ben-Porath, Y. S., & Tellegen, A. (2008). *Minnesota Multiphasic Personality Inventory-2-Restructured Form: Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Bersin, J. (2011). *Strategic Human Resources and Talent Management: Predictions for 2012. Driving Organizational Performance amidst an Imbalanced Global Workforce.* Bersin & Associates research report. Retrieved 11-12-2011 from http://www.bersin.com/

Binet, A., & Henri, V. (1895). La psychologie individuelle. *L'Annee Psychologique, 2*, 411-465.

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89*, 150-157.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11*, 515-524.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Bouwmeester, S., & Sijtsma, K. (2006). Constructing a transitive reasoning test for 6- to 13-year-old children. *European Journal of Psychological Assessment, 22*, 225-232.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910-919.

Briggs, S. R. (1989). The optimal level of measurement for personality constructs. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp.246-260). New York: Springer-Verlag.

Briggs, S. R. (1992). Assessing the five-factor model of personality description. *Journal of Personality, 60*, 253-293.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2*, 171-183.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460-502.

Burke, E. (2010). Research findings on LOFT-based model. *ATP 2010, Innovations in Testing. Session: Verification Testing Models for Unproctored Assessment: Issues, Research & Practice,* Orlando, Florida.

Burke, E. (2011). *Talent Management. Special report: Assessment and Measurement. How's Your Talent Doing?* MediaTec Publishing Inc. Retrieved 10-12-2011 from http://talentmgt.com/articles/view/assessments-and-measurement

Centraal Bureau voor de Statistiek (CBS) [Dutch Central Bureau of Statistics] (2000, November/December). *Standaarddefinitie allochtonen* [Standard definition of immigrants]. Retrieved from http://www.cbs.nl/NR/rdonlyres/26785779-AAFE-4B39-AD07-59F34DCD44C8/0/index1119.pdf

Chernyshenko, O.S., & Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M.D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22,* 105 - 127.

Chernyshenko, O. S., Stark, S., & Drasgow, F. (2010). Individual differences, their measurement, and validity. In S. Zedeck (Ed.). *APA Handbook of Industrial and Organizational Psychology, Vol. 2: Selecting and Developing Members for the Organization (pp. 117-151)*. Washington, DC: American Psychological Association.

Cheung, F. M., Leung, K., Zhang, J. X., Sun, H. F., Gan, Y. Q., Song, W. Z., & Xie, D. (2001). Indigenous Chinese personality constructs: Is the five-factor model complete? *Journal of Cross-Cultural Psychology, 32*, 407-433.

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 19*, 125-136.

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? no. *Journal of Research in Personality, 40*, 359-376.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Clark, L. A. (1993). *Schedule for Nonadaptive and Adaptive Personality (SNAP). Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1998). Trait theories of personality. In D. F. Barone, M. Hersen, V. B. Van Hasselt, D. F. Barone, M. Hersen & V. B. Van Hasselt (Eds.), *Advanced personality*. (pp. 103-121). New York, NY US: Plenum Press.

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on the DAS and KAIT. In S. E. Embretson, & S. L. Hershberger (Eds.). *The new rules of measurement: What every psychologist should know* (pp. 37-63). Mahwah, NJ: Erlbaum.

De Raad, B. (1992). The replicability of the big five personality dimensions in three word-classes of the dutch language. *European Journal of Personality, 6*, 15-29.

Diesfeldt, H. F. A. (2004). Executive functioning in psychogeriatric patients: Scalability and construct validity of the Behavioral Dyscontrol Scale (BDS). *International Journal of Geriatric Psychiatry, 19*, 1065-1073.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Do, B., Shepherd, W., & Drasgow, F. (2005). *Measurement equivalence across proctored and unproctored administration modes of web -based measures*. Paper Presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. Journal of Applied Psychology, 91, 40-57.

Egberink, I. J. L., & Meijer, R. R. (2011). An item response theory analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. *Assessment, 18*, 201-212.

Egberink, I. J. L., & Meijer, R. R. (2012). Het nut van de item respons theorie bij de constructie en evaluatie van niet-cognitieve instrumenten voor selectie en assessment binnen organisaties. *Gedrag & Organisatie, 25*, 87-107.

Egberink, I. J. L., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44*, 232-244.

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*, 386-395.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105-120.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10,* 295-317.

Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research, 18*, 23-36.

Forbey, J. D., Ben-Porath, Y., & Arbisi, P. A. (2011). The MMPI–2 computerized adaptive version (MMPI–2–CA) in a veterans administration medical outpatient facility. *Psychological Assessment*, doi:10.1037/a0026509.

Foster, D. (2009). Secure, online, high-stakes testing: Science fiction or business reality? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 31-34.

Foster, D., Maynes, D., & Hunt, B. (2008). Using data forensic methods to detect cheating. In C. L. Wild, & R. Ramaswamy (Eds.), *Improving testing: Applying process tools and techniques to assure quality* (pp. 305-322). Mahwah, NJ: Lawrence Erlbaum Associates.

Foxcroft, C. D., & Davies, C. (2006). Taking ownership of the ITC's guidelines for computer-based and Internet-delivered testing: A South African application. *International Journal of Testing, 6*, 173-180.

Furnham, A., & Fudge, C. (2008). The Five Factor Model of personality and sales performance. *Journal of Individual Differences, 29*, 11-16.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., & Grochocinski, V. J. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 57*, 361-368.

Gibby, R. (2010). Example global UIT program with verification. *ATP 2010, Innovations in Testing. Session: Verification Testing Models for Unproctored Assessment: Issues, Research & Practice,* Orlando, Florida.

Gnambs, T., & Batinic, B. (2011). Evaluation of measurement precision with Rasch-type models: The case of the short generalized opinion leadership scale. *Personality and Individual Differences, 50*, 53-58.

Goldberg, L. R. (1990). An alternative 'description of personality': The big-five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, K. Widaman, D. C. Funder, R. D. Parke, . . . K. Widaman (Eds.), *Studying lives through time: Personality and development.* (pp. 169-188). Washington, DC US: American Psychological Association.

Gough, H. G. (1957). *Manual for the California Psychological Inventory*™. Mountain View, CA: CPP, Inc.

Gough, H. G., & Heilbrun, A. B. (1980). *The adjective Check List manual 1980 edition.* Palo Alto, CA: Consulting Psychologists Press.

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89*, 167-187.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18*, 351-364.

Guthridge, M., Komm, A.B., Lawson, E. (2008), "Making talent a strategic priority", *The McKinsey Quarterly*, No.1.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 209-227.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press.

Heller, D., Ferris, D. L., Brown, D., & Watson, D. (2009). The influence of work personality on job satisfaction: Incremental validity and mediation effects. *Journal of Personality, 77*, 1051-1084.

Hoekstra, H. A., Ormel, J., & de Fruyt, F. (2007). *NEO-PI-R en NEO-FFI persoonlijkheidsvragenlijsten: Handleiding* [Manual for the Dutch version of the NEO-PI-R]. Amsterdam: Hogrefe.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8*, 149-162.

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63*, 146-163.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270-1285.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2001). Toepassing van een computergestuurde adaptieve testprocedure op persoonlijkheidsdata [Application of a computerized adaptive test procedure on personality data]. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 56*, 119-133.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2005). A randomized experiment to compare conventional, computerized, and computerized adaptive administration of ordinal polytomous attitude items. *Applied Psychological Measurement, 29*, 159-183.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. *Applied Psychological Measurement, 31*, 412-429.

Hough, L. M. (1992). The 'big five' personality variables—construct confusion: Description versus prediction. *Human Performance, 5*, 139-155.

Hough, L. M. (1998). Effects of intentional response distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209-244.

Hough, L. M., & Dilchert, S. (2007). *Inventors, innovators, and their leaders: Selecting for Conscientiousness will keep you "inside the box"*. Paper presented at SIOP's 3rd Leading Edge Consortium: Enabling Innovation in Organizations, Kansas City, MO.

Hough LM, Ones DS, Viswesvaran C. (1998, April). *Personality correlates of managerial performance constructs*. In Page R (Chair). Personality determinants of managerial

potential, performance, progression and ascendancy. Symposium conducted at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future - remembering the past. *Annual Review of Psychology, 51*, 631-664.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial–organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 272-290.

Howard, P. J., & Howard, M. J. (2001). *Professional manual for the Workplace Big Five profile (WB5P).* Charlotte, NC: Centacs.

Hunthausen, J. M., Truxillo, D. M., Bauer, T. N., & Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity. *Journal of Applied Psychology, 88*, 545-551.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. Journal of Applied Psychology, 85, 869-879.

Interapy (n.d.). http://www.interapy.nl/

International Test Commission (ITC). (2005). *International guidelines on computer-based and internet-delivered testing.* downloaded electronically on 2011/09/17 from www.intestcom.org/itc_projects.htm.

Jackson, D. N. (1976). *Jackson personality inventory manual.* Port Huron MI: Research Psychologists Press.

Jackson, D. N. (1997). *Jackson personality inventory manual revised.* Port Huron MI: Research Psychologists Press.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*, 843-877.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues (6th ed.).* Belmont, CA: Thomson Wadsworth.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory, *Multivariate Behavioral Research, 39*, 329-358.

Lei, H., & Dai, X. (2011). Application of computerized adaptive testing to eysenck personality questionnaire. *Chinese Journal of Clinical Psychology, 19*, 306-308.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology, 84*, 817-824.

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*, 268-279.

Lievens, F., & de Soete, B. (2011). Instrumenten om personeel te selecteren in de 21ste eeuw: Onderzoek en praktijk. *Gedrag & Organisatie, 24*, 18-42.

Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578-595.

MacCann, C., Ziegler, M. & Roberts, R.D. (2011). Faking in Personality Assessment. In M. Ziegler, C. MacCann, and R. D. Roberts (Eds.) *New perspectives on faking in personality assessment.* New York: Oxford University Press.

Maij-de Meij, A. M., Schakel, L., Smid, N., Verstappen, N., & Jaganjac, A. (2008). *Connector Ability 1.1, Professional Manual.* Utrecht: PiCompany B.V.

Maij-de Meij, A., & Schakel, L. (2011). *Improving a short CAT verification test: Simulation versus practice.* 2011 IACAT Conference, Pacific Groove, California.

Makransky, G., & Glas, C. A. W. (2011). Unproctored internet test verification: Using adaptive confirmation testing. *Organizational Research Methods, 14*, 608-630.

McCrae, R. R. (.)., & Allik, J. (2002). In McCrae R. R., Allik J. (Eds.), *The five-factor model of personality across cultures*. New York, NY US: Kluwer Academic/Plenum Publishers.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

McHorney, C.A. & Cohen, A.S. (2000). Equating health status measures with item response theory: Illustrations with functional status items. *Medical Care, 38*, 43-59.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388.

Meade, A.W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*, 728-743.

Meijer, R. R. (2010). A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale? *Personality and Individual Differences, 48*, 502-503.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368.

Meijer, R. R., & Egberink, I. J. L. (in press). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment, 90*, 227-238.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*, 187-194.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2010). Detecting person misfit in adaptive testing. In W. J. van der Linden & C. A. W. Glas: *Elements of adaptive testing* (pp. 315-329). New York: Springer.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology, 55*, 675-680.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement, 69,* 613-635.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5.0 for windows. User's manual*. Groningen, The Netherlands: ProGAMMA.

Moorer, P., Suurmeijer, Th. P. B. M., Foets, M., & Molenaar, I. W. (2001). Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research, 10*, 637-645.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). 'Reconsidering the use of personality tests in personnel selection contexts,'. *Personnel Psychology, 60*, 683-729.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). 'Are we getting fooled again? coming to terms with limitations in the use of personality tests for personnel selection,'. *Personnel Psychology, 60*, 1029-1049.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59-71.

Myers, I. B. (1962). *Manual the myers-briggs type indicator*. Palo Alto CA: Consulting Psychologists Press.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the internet: New problems, old issues. *American Psychologist*, 59, 150-162.

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw Hill, Inc.

Nye, C. D., Do, B., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment, 16*, 112-120.

Onderzoek bij testkandidaten. (n.d.). http://www.psynip.nl/website/wat-doet-het-nip/tests/testkandidaten/onderzoek-bij-testkandidaten

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

Oud, A. N. M., Bloemers, W., & Reitz, E. (2009). Loont bedrog? Effecten van fraude bij een online intelligentietest zonder supervisie. *Gedrag & Organisatie, 22*, 200-213.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74*, 538-556.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Sover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment, 18*, 263-283.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality,

biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733-752.

Potosky, D., & Bobko, P. (1997). Computer versus paper-and-pencil administration mode and response distortion in noncognitive selection tests. *Journal of Applied Psychology, 82*, 293-299.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education, 23*, 266-285.

Raju, N. (1998). *DFITP4: A Fortran program for calculating DIF/DTF* [Computer software]. Chicago: Illinois Institute of Technology.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, *7*, 347-364.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93-103.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754-775.

Rivas, T., Bersabé, R., & Berrocal, C. (2005). Application of the double monotonicity model to polytomous items: Scalability of the Beck Depression Items on subjects with eating disorders. *European Journal of Psychological Assessment, 21*, 1-10.

Roberts, B. W. (2006). Personality development and organizational behavior (Chapter 1, pp 1-41). In B. M. Staw (Ed.). *Research on Organizational Behavior*. Elsevier Science/JAI Press.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

Roorda, L. D., Roebroeck, M. E., van Tilburg, T., Molenaar, I. W., Lankhorst, G. J., & Bouter, L. M. (2005). Measuring activity limitations in walking: Development

of a hierarchical scale for patients with lower-extremity disorders who live at home. *Archives of Physical Medicine and Rehabilitation*, *86*, 2277-2283.

Rudner, L. M. (2010). Implementing the graduate management admission test computerized adaotiev test. In: W. J. van der Linden & C.A.W. Glas: *Elements of Adaptive Testing* (pp. 151-165).

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assessees' perceptions and reactions. *International Journal of Selection and Assessment, 11*, 194-205.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Saucier, G. (2003). An alternative multi-language structure of personality attributes. *European Journal of Personality*, *17*, 179-205.

Schakel, L., Smid, N. G., & Jaganjac, A. (2007). *Workplace Big Five professional manual.* Utrecht, The Netherlands: PiCompany B.V.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments & Computers, 29*, 274-279.

Schmitt, D. P., Allik, J., Mccrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology, 38, 173-212.

Schreurs, B., Derous, E., Proost, K., & De Witte, K. (2010). The relation between selection expectations, perceptions and organizational attraction: A test of competing models. *International Journal of Selection and Assessment, 18*, 447-452.

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment, 13*, 442-453.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63*, 183-200.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79-105.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences, 50*, 31-37.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (SNAP). *Psychological Assessment, 17,* 28-43.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45*, 553-573.

Sodano, S. M., & Tracey, T. J. G. (2011). A brief inventory of interpersonal problems-circumplex using nonparametric item response theory: Introducing the IIP-C-IRT. *Journal of Personality Assessment, 93*, 62-75.

Stark, S. Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*, 497-508.

Stark, S., & Chernyshenko, O. S. (2011). Computerized Adaptive Testing with the Zinnes and Griggs Pairwise Preference Ideal Point Model. *International Journal of Testing, 11*, 231-247.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? Journal of Applied Psychology, 89, 497-508.

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior, 24*, 1216-1228.

Tendeiro, J. N., & Meijer, R. R. (in press). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*.

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (in press). Using Cumulative Sum statistics to detect inconsistencies in unproctored Internet testing. *Educational and Psychological Measurement.*

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to morgeson, campion, dipboye, hollenbeck, murphy, and schmitt (2007). *Personnel Psychology, 60*, 967-993.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Thissen, D. (2001). *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning* [Computer software]. Chapel Hill: University of North Carolina at Chapel Hill.

Thissen, D. (2003). *Multilog for Windows (Version 7.0)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in tracelines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.147-172). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Tippins, N. T. (2009a). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 2-10.

Tippins, N. T. (2009b). Where is the unproctored Internet testing train headed now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 69-76.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59*, 189-225.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87*, 1020-1031.

van der Linden, W. J., & Glas, C. A. W. (Eds.) (2000). *Computerized adaptive testing: Theory and practice*. Norwell, MA: Kluwer.

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1-19.

van der Linden, W. J.,& Glas, C. A.W. (2010). *Elements of Adaptive Testing*. New York: Springer:

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-218.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detection of person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston: Kluwer-Nijhof.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional constrained adaptive testing. *Psychometrika, 67*, 575-588.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Waller, D. (1993, April 12). A Tour Through 'Hell Week': A Newsweek correspondent takes the CIA spy tests. *Newsweek*, 33.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 185-217). Mahwah, NJ: Lawrence Erlbaum.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption Scale. *Journal of Personality and Social Psychology, 57,* 1051-1058.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Quality of Life Research, 16*, 143-155.

Watson, R., Deary, I., & Austin, E. (2007). Are personality trait items reliably more or less 'difficult'? Mokken scaling of the NEO-FFI. *Personality and Individual Differences, 43*, 1460-1469.

Watson, R., Roberts, B., Gow, A., & Deary, I. (2008). A hierarchy of items within Eysenck's EPI. *Personality and Individual Differences, 45*, 333-335.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance, 17*, 433-461.

Weiner, J. A. (2010). Overview of models, considerations, value. *ATP 2010, Innovations in Testing. Session: Verification Testing Models for Unproctored Assessment: Issues, Research & Practice,* Orlando, Florida.

Weiner, J. A., Knapp, D. J., & Hogan, J. B. (2011). Unproctored internet testing: Issues & considerations for different assessment contexts. *ATP 2011, Innovations in Testing,* Phoenix, Arizona.

Weiner, J. A., & Ruch, W. W. (2006). Effects of cheating in unproctored internet based testing: A monte carlo study. *21st Annual SIOP Conference.* Dallas.

Weiss, D. J. (2004). Computerized Adaptive Testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37,* 70-84.

Wiechmann, D., & Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment*, 11, 215-229.

Wilkerson, J. M., Nagao, D. H., & Martin, C. L. (2002). Socially desirable responding in computerized questionnaires: When questionnaire purpose matters more than the mode. *Journal of Applied Social Psychology, 32*, 544-559.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

Summary

Historically, early psychologists working in a business setting influenced personnel selection by relying on the scientific methodology of experimental psychology grounded in the measurement of individual differences of empirically verifiable observations. This research depended on progress in both measurement and statistics and reflected a pragmatic approach. This same pragmatic approach is seen today with the development and the use of computer-based testing. In recent years, computer-based testing has become popular in human resource management (HRM) practice, especially for the administration, scoring, and reporting of test scores in personnel selection and in career development settings. Also the use of the Internet in combination with proctored and unproctored testing is increasing. In contrast to the early days of scientific personnel selection, the use of the scientific methodology is often ignored. This thesis tries to fill this gap by applying psychometric models and procedures for the development of an online computer-based Big Five instrument for the workplace, the Reflector Big Five Personality (RBFP). Psychometric research that evaluates the quality of this instrument is discussed and methods that can help to obtain information about the validity of scores that are obtained in an unproctored setting are discussed and compared. Because the validity of scores is especially a problem in maximum performance testing, a cognitive computer-based test is used for this latter research.

In Chapter 1, personality testing in the workplace is introduced. The usefulness of personality testing is discussed and new developments and recent changes of personality testing within HRM as a result of computer use and the intense use of the Internet are sketched.

In Chapter 2, the theoretical and psychometrical background of the RBFP questionnaire is discussed. This is done, because in Chapters 3 through 5 different psychometric methods, based on IRT, are applied to this instrument. In Chapter 3 differential item and test functioning of the RBFP is investigated using different types of effect size measures, in Chapter 4 a computerized adaptive version of the RBFP is developed and discussed, and in Chapter 5 the property of invariant item ordering (IIO) is investigated for the RBFP. The RBFP is an online-administered computer-based Big Five instrument, therefore in Chapter 2 the Big Five model and its use within HRM are discussed first. Second, the development of the RBFP and

some research studies regarding its psychometric quality are described. Finally, the online administration and reporting process are discussed.

In Chapter 3, it is investigated whether the items and the subtest of the RBFP have similar psychometric properties in different populations. In this chapter differential functioning of the RBFP is investigated in two contexts, a selection context and a career development context. First, scaling results are compared for the selection and development context. Second, differential item and test functioning are investigated using a likelihood ratio approach and using different effect size measures. Results showed that the scalability was lower in the selection context than in the developmental context, but that differential test functioning was of no practical importance.

In Chapter 4, the usefulness of computerized adaptive testing (CAT) for personality in a real life personnel selection context is investigated. A sample of candidates completed the CAT as part of a career development procedure. Results showed that CAT resulted in a reduction of approximately 50% of the items administered and administration time, whereas high correlations were found between CAT and full scale scores. However, the item pool was not very suited to discriminate candidates with moderate to high values on the investigated personality traits. Item administration order demonstrated variability across candidates.

In Chapter 5, it is investigated whether subsets of items from the RBFP have the property of invariant item ordering (IIO). This property may be used to select items for short questionnaires and may help to obtain insight in the general quality of the individual items. Because IIO research is an unexploited field in test construction and test evaluation, the usefulness of a recently proposed method is proposed. Many items of the RBFP did not comply to this property.

Unproctored Internet testing (UIT) is becoming more popular in personnel recruitment and selection. A drawback of UIT procedures is that cheating is easy, and, therefore, a proctored test is often administered after an UIT procedure. For a particular person, cheating may result in large differences between inconsistent test scores across different test modes. To detect inconsistent test scores across unproctored and proctored test scores, in Chapter 6 different statistical methods to detect inconsistent test scores are discussed. Furthermore, a new methodology based on the cumulative sum methodology was applied. In this new methodology latent trait estimates on the unproctored test are used to investigate the likelihood of an item score pattern on the proctored test. The idea behind this procedure is that when candidates are cheating, that is, get help from a more able person, their estimated trait value on the UIT is not a good indication of their true trait level and

large differences between estimated UIT and UT trait values when a person is retested are expected. The usefulness of the CUSUM is illustrated and the unique contribution of the CUSUM to existing procedures is discussed.

Samenvatting (Summary in Dutch)

Binnen human resource management spelen psychologische instrumenten zoals persoonlijkheidsmetingen of intelligentietests van oudsher een belangrijke rol in het selecteren van sollicitanten voor vacatures of voor het geven van loopbaanadvies. Het gebruik van internettoepassingen maakt het mogelijk om het selectieproces of loopbaanadvisering deels online te doorlopen. Zo wordt het bijvoorbeeld steeds gebruikelijker dat een sollicitant bij zijn of haar sollicitatie psychologische tests geheel via internet invult. Dit betekent dat het gebruik van de instrumenten verandert en dat andere eisen aan de instrumenten worden gesteld. Zo is bij het invullen van vragenlijsten en het maken van tests soms geen toezicht meer, worden eisen aan de maximale lengte van de vragenlijsten en tests gesteld en kunnen testgegevens die in databanken zijn opgeslagen gemakkelijk aangewend worden voor onderzoek naar geschiktheid voor verschillende functies.

Het goed in kaart kunnen brengen van individuele verschillen tussen personen is in sterke mate afhankelijk van de kwaliteit van de gebruikte instrumenten. Om de psychometrische kwaliteit van tests en vragenlijsten in kaart te brengen speelt traditioneel de klassieke testtheorie (KTT) een grote rol. Daarnaast is sinds de jaren '50 van de vorige eeuw de item respons theorie ontwikkeld en IRT is op sommige terreinen de standaardmethode geworden om testgegevens te analyseren. Hoewel we de afgelopen jaren een toename zien van het gebruik van IRT binnen het psychologisch meten, getuige ook de opname van allerlei IRT criteria waaraan psychologische tests dienen te voldoen in de COTAN handleiding, wordt er nog weinig gebruik gemaakt van deze technieken in, met name, het niet-cognitieve domein (bijv. persoonlijkheid, interesse, attitude).

Historisch gezien is de selectiepsychologie altijd gekenmerkt door een grote mate van pragmatisme. Hoewel de eerste psychologen die werkzaam waren in een bedrijfscontext de wetenschappelijke methode van het in kaart brengen van individuele verschillen baseerden op statistische en psychometrische methoden, zien we dat bij de opkomst en het gebruik van de computer en internettoepassingen binnen HRM hetzelfde pragmatisme de overhand voert, maar vaak wordt dit niet ondersteund door gebruik te maken van de nieuwe ontwikkelingen in de psychometrie. Het doel van deze these is om deze leemte te vullen.

In Hoofdstuk 1 wordt de ontwikkeling geschetst van de persoonlijkheidsmeting in de bedrijfscontext met de nadruk op het hedendaags gebruik van de computer en het internet.

In Hoofdstuk 2 wordt de theoretische en psychometrische ontwikkeling besproken van de Reflector Big Five Persoonlijkheidsvragenlijst (RBFP). De RBFP is een vragenlijst die online wordt afgenomen en speciaal is ontwikkeld voor de bedrijfscontext. Zowel de scoring, de manier waarop de vragenlijst wordt afgenomen als de rapportage worden besproken.

In Hoofdstuk 3 wordt onderzocht of de items en de subtests dezelfde psychometrische eigenschappen hebben in verschillende populaties. Item- en testzuiverheid wordt onderzocht voor twee verschillende contexten: een selectiecontext en een ontwikkelcontext. Eerst wordt gekeken of de schalingseigenschappen hetzelfde zijn in de twee verschillende contexten. Daarna, worden een likelihood ratio test en verschillende maten om de effectgrootte te meten gebruikt om zowel item- als testonzuiverheid te bepalen. Resultaten wijzen erop dat de schaalbaarheid geringer was in de selectiecontext, maar dat er geen sprake is van testonzuiverheid.

In Hoofdstuk 4 wordt het gebruik van een adaptieve testprocedure voor de RBFP onderzocht. Resultaten laten zien dat deze procedure leidt tot een reductie van 50% van de aangeboden items en een reductie van 50% van de tijd die men kwijt is aan het invullen van de items. Wat echter ook opviel was dat de item pool niet erg geschikt was om personen met een gemiddelde tot een hoge trek score van elkaar te onderscheiden. Verder bleek de ordening van de aangeboden items te verschillen per kandidaat.

In Hoofdstuk 5 wordt de eigenschap van invariante item ordening onderzocht bij de RBFP. Hoewel deze eigenschap niet noodzakelijk is voor de ordening van personen naar hun latente trek score, kan het een nuttige methode zijn om, bijvoorbeeld, items te selecteren uit een item pool die een breed aantal kenmerken van de te meten trek meet. In Hoofdstuk 5 wordt geconcludeerd dat er de nodige schendingen zijn wat betreft IIO bij de items van de RBFP en dat een aantal items replicaties van elkaar zijn.

Het gebruik van unproctored tests (dat wil zeggen testafnames zonder toezicht) wordt steeds populairder. Vaak worden deze unproctored tests gevolgd door een proctored test (dat wil zeggen een testafname met toezicht) om te controleren of een kandidaat de unproctored test zelf heeft gemaakt of dat er bedrog in het spel is. Om te controleren of een kandidaat eerlijk is geweest bij de beantwoording van de vragen, zijn verschillende psychometrische methoden beschikbaar. In Hoofdstuk 6

wordt onderzoek gedaan naar verschillende methoden om latente trek schattingen van proctored en unproctored tests te vergelijken. Ook wordt een nieuwe methode toegepast die gebaseerd is op de 'Cumulative Sum Procedure'. Het idee bij deze methode is dat een item score patroon op de proctored test onwaarschijnlijk is wanneer we dit patroon analyseren met de latente trek waarde van kandidaat die bedrog heeft gepleegd op de unproctored test.

Dankwoord (Acknowledgements)

In mijn werk bij PiCompany heb ik veel tests en vragenlijsten mogen ontwikkelen. Kwaliteit en praktische toepasbaarheid binnen HRM zijn altijd belangrijke criteria geweest bij het ontwikkelen van deze instrumenten. Aan de basis hiervan stond de toepassing van een goede mix van psychologische kennis, psychometrische methoden en toepassingen van nieuwe IT-technologie. Een samenwerkingsverband met de Universiteit Twente bij het ontwikkelen van een adaptieve cognitieve capaciteitentest bracht mij in aanraking met de item respons theorie (IRT). Deze techniek bleek vele mogelijkheden te hebben om, gezien de opkomst van online testafnames, nieuwe vragen uit de HRM praktijk te kunnen beantwoorden. De kennismaking met deze techniek in combinatie met mijn ervaringen als testontwikkelaar gaven aanleiding tot dit proefschrift.

Een proefschrift kan niet tot stand komen zonder de medewerking en ondersteuning van vele mensen. Hen wil ik hierbij danken.

Allereerst wil ik mijn promotor Rob Meijer bedanken. Rob, je wetenschappelijke vakmanschap en uitgebreide kennis van methoden en technieken waren een grote inspiratie voor mij. Je hebt er voor gezorgd dat het tempo in mijn promotie bleef. Je gaf vertrouwen dat ik mijn promotie-activiteiten naast mijn werk bij PiCompany kon doen. Daarnaast heb je me gewezen op zaken die voor mij in mijn werk vanzelfsprekend waren, maar die ook wetenschappelijke relevantie hadden. Omgekeerd vond ik het stimulerend om nieuwe ideeën over online testgebruik en toekomstige ontwikkelingen te bespreken.

Als tweede wil ik mijn co-promotor Iris Egberink bedanken. Iris, tijdens jouw promotieonderzoek hebben we al samengewerkt. De voortzetting van deze samenwerking tijdens mijn promotieonderzoek heb ik als zeer plezierig ervaren. Zonder jouw kritische blik, consciëntieusheid en wetenschappelijke stofkam zou dit proefschrift niet zijn zoals het nu is.

Rob en Iris, samen hebben jullie mijn praktijkgerichte focus verrijkt met een academische blik. Wat ik vanzelfsprekend vond vanuit de praktijk, stelden jullie ter discussie. Maar omgekeerd gold dit ook. Mijn ervaringen uit de dagelijkse praktijk, waren voor jullie een aanvulling op jullie wetenschappelijke focus. Met de vooruitzichten van leuke besprekingen en een open en gezellige sfeer, was de twee uur durende reis naar Groningen nooit een belasting. Erg bedankt voor jullie steun bij de totstandkoming van dit proefschrift.

Graag wil ik ook PiCompany bedanken voor de bijdrage aan dit proefschrift. PiCompany heeft mij ondersteund door testdata en tijd beschikbaar te stellen voor deze promotie. Mijn voormalig directeur Martijn van der Woude heeft mij gestimuleerd om mijn kennis en ervaring binnen testontwikkeling te gebruiken voor een promotietraject. Later werd ik hierin nog intensiever gemotiveerd door Pieter van Hoogstraten. Hem wil ik met name bedanken voor zijn ondersteuning en het belang dat hij hechtte aan mijn promotie.

Verder bedank ik graag de leden van de beoordelingscommissie, Klaas Sijtsma, Karin Sanders en Hans Hoekstra, voor het zorgvuldig lezen van mijn proefschrift en de aanwezigheid tijdens mijn promotie. Tevens bedank ik de overige leden van de promotiecommissie voor hun aanwezigheid.

Van de Rijksuniversiteit Groningen wil ik ook Jorge Tendeiro bedanken. Zijn kennis over CUSUM-grafieken en de analyses van de benodigde data zijn van grote waarde geweest bij het detecteren van afwijkende patronen in testafnames.

Daarnaast wil ik mijn directe collega's bij PiCompany bedanken. Speciaal wil ik daarbij Nico Smid vermelden. Met hem heb ik een groot gedeelte van mijn tijd bij PiCompany op een zeer plezierige manier samengewerkt. Zijn passie voor psychometrie, zijn conceptuele denkkracht en zijn pragmatische aanpak zal een blijvende inspiratiebron voor mij blijven. Daarnaast wil ook ik de bijdrage van Annette Maij – de Meij niet onvermeld laten. Haar grote IRT-kennis en ervaring was een bron waar ik altijd op terug kon vallen. Ook Noortje Verstappen en Carin Bossink wil ik bedanken voor hun werk aan de instrumenten die ik in mijn onderzoek heb gebruikt. Maar bovenal wil ik jullie alle vier bedanken voor de plezierige samenwerking.

Ook Pierce en Jane Howard van Centacs wil ik niet onvermeld laten. Hun pionierswerk over het toegankelijk maken van persoonlijkheidsmetingen in een werkcontext leidde uiteindelijk tot de Reflector Big Five vragenlijst waar een deel van het onderzoek in dit proefschrift op is gebaseerd.

Buiten het werk wil ik eerst mijn partner Brunhilde bedanken voor haar interesse en het geduld dat zij heeft opgebracht. Regelmatig heb ik avonden en weekenden besteed aan het gestaag verder werken aan het proefschrift. Daarnaast wil ik natuurlijk ook mijn moeder, familie, schoonfamilie en vrienden bedanken voor hun mentale ondersteuning, bemoedigende woorden en interesse in de voortgang van mijn promotie.